# A Logic of Universal Causation

Hudson Turner

*Department of Computer Science, University of Minnesota at Duluth, Duluth, Minnesota 55812, USA, e-mail:* `hudson@d.umn.edu`

**Abstract**

For many commonsense reasoning tasks associated with action domains, only a relatively simple kind of causal knowledge is required—knowledge of the conditions under which facts are caused. This note introduces a modal nonmonotonic logic for representing causal knowledge of this kind, relates it to other nonmonotonic formalisms, and shows that a variety of causal theories of action can be expressed in it, including the recently proposed causal action theories of Lin. The new logic extends the causal theories formalism of McCain and Turner, and provides a more adequate semantic account of it. A useful subset of the logic has a concise translation into classical propositional logic, and so can be used for automated planning and reasoning about action. A larger subset is closely related to logic programming under the answer set semantics, yielding another approach to automated reasoning.

*Keywords:* Reasoning about actions; Causal logic; Nonmonotonic logic

## 1 Introduction

This note introduces a modal nonmonotonic logic of "universal causation," called UCL, designed for formalizing commonsense knowledge about actions. UCL extends the recently introduced causal theories formalism of McCain and Turner [39], which shares its underlying motivations. The fundamental distinction in UCL—between propositions that have a cause and propositions that (merely) obtain—is expressed by means of the modal operator $\mathsf{C}$, read as "caused." For example, one can write

$$\phi \supset \mathsf{C}\psi$$

to say that $\psi$ is caused whenever $\phi$ obtains. These simple linguistic resources make it possible for a UCL theory to express the conditions under which facts are caused. It is in this sense that UCL is a logic of causation.

As usual for nonmonotonic logics, the main semantic definition in UCL—of a "causally explained" interpretation—is given by a fixpoint condition. Intuitively, an interpretation is causally explained by a UCL theory $T$ if it represents the facts true in a world that is "causally possible" according to $T$. The focus in UCL on causally possible worlds is motivated by the following pair of observations.

- Knowledge of the causally possible worlds is sufficient for many commonsense reasoning tasks associated with action domains, such as prediction and planning.
- In order to determine the causally possible worlds, it is sufficient to know the conditions under which facts are caused.

The first observation suggests that UCL can be useful. The second observation helps explain why UCL can be simple: it formalizes causal knowledge of a relatively simple kind, and does not attempt the notoriously difficult task of formalizing causal relations of the form "$\phi$ causes $\psi$." Happily, one can describe and reason about the conditions under which facts are caused without settling the question of what causes what.

In UCL, the notion of a causally possible world is made precise on the basis of the following pair of assumptions.

- In a causally possible world, every fact that is caused obtains.
- In a causally possible world, every fact that obtains is caused.

The first assumption is unremarkable. The second is not. As in [39], we call it the *principle of universal causation*. This simplifying assumption is the key to the main semantic definition of the logic, which is therefore named for it. We take these two assumptions together to define what it is for a world to be causally possible: what obtains in the world is exactly what is caused in it. Accordingly, the main semantic definition in UCL says that an interpretation $I$ is causally explained by a UCL theory $T$ if what is true in $I$ is exactly what is caused in $I$ according to $T$.

The principle of universal causation is easily relaxed in practice. For instance, when describing action domains, we generally have little or nothing to say about the "actual" conditions under which facts in the initial situation are caused. Instead, our UCL action theories typically stipulate that facts in the initial situation are caused. Such stipulations are straightforward in UCL, where one can say that $\phi$ is caused whenever it is true simply by writing

$$\phi \supset \mathsf{C}\phi .$$

In the same way, we typically stipulate that facts about the occurrence and non-occurrence of actions are caused.

More interesting are those facts that, roughly speaking, are true simply because they were true before and haven't been made false since. It is facts of this kind that give rise to the frame problem [42]. Solutions to the frame problem typically appeal to the "commonsense law of inertia," according to which the value of an inertial fluent persists unless it is caused to change. The principle of universal causation makes possible a simple, robust encoding of the commonsense law of inertia. Take $f_t$ to stand for the proposition that a fluent $f$ holds at a time $t$. One can write

$$f_t \wedge f_{t+1} \supset \mathsf{C} f_{t+1} \tag{1}$$

to stipulate that $f$ is caused at time $t+1$ whenever it persists from time $t$ to time $t+1$. Thus, axioms of form (1) can, in effect, suspend the principle of universal causation with respect to persistent inertial fluents. Of course, universal causation still requires that if $f$ does not persist, the new value of $f$ must be caused. That is, the UCL theory must describe conditions sufficient for it to be caused. In this way, inertia axioms of form (1) interact with the principle of universal causation to solve the frame problem, guaranteeing that inertial fluents persist unless they are caused to change.

Typical features of action domain descriptions are easily expressed in UCL. Here are a few examples based on the infamous Yale Shooting domain [23] (as extended by Baker [1]). One can write

$$Shoot_t \supset \mathsf{C}\neg Alive_{t+1} \wedge \mathsf{C}\neg Loaded_{t+1} \tag{2}$$

to describe the direct effects of shooting: whenever *Shoot* occurs, both $\neg Alive$ and $\neg Loaded$ are caused to hold subsequently. One can write

$$Shoot_t \supset Loaded_t$$

to express a precondition of the shoot action: shoot can occur only when the gun is loaded. To say that Fred is caused to be not walking whenever he is caused to be not alive, one can write, for instance,

$$\mathsf{C}\neg Alive_t \supset \mathsf{C}\neg Walking_t\,. \tag{3}$$

From (2) and (3) it follows that whenever *Shoot* occurs, $\neg Walking$ is caused to hold subsequently. Traditionally, $\neg Walking$ is referred to as an "indirect effect" or "ramification" of the action *Shoot*.

In accordance with common sense, (3) does not imply $\mathsf{C}\, Walking_t \supset \mathsf{C} Alive_t$. Intuitively, you cannot bring Fred back to life by getting him to walk. Instead, he simply can't walk unless he's alive. That is, in any causally possible world, Fred is alive if he is walking. Accordingly, if (3) is an axiom of a UCL theory $T$, then $Walking_t \supset Alive_t$ is true in every interpretation causally explained by $T$.

UCL differs in fundamental ways from nonmonotonic formalisms such as default logic [47] and autoepistemic logic [43]. First, UCL is not motivated by the problem of general default reasoning and knowledge representation. It is designed for a more specific purpose. Second, in UCL one describes the conditions under which facts are caused, rather than the conditions under which facts are believed or known. Third, the fixpoint condition in UCL characterizes complete worlds, in the form of classical interpretations, rather than incomplete, logically closed belief sets. Nonetheless, we will see that UCL is closely related to default logic, in the special case when we consider only the "complete," consistent extensions of default theories. We will also discuss a rather striking similarity between the main semantic definitions of UCL and autoepistemic logic.

The syntax and some of the motivations of UCL are anticipated in a more ambitious formalism introduced by Geffner [11–13]. Geffner employs a modal language with a single modal operator $\mathsf{C}$, read as "explained," and defines "default theories which explicitly accomodate a distinction between 'explained' and 'unexplained' propositions" [12]. His proposal is meant to enhance "the appeal of preferential entailment as a unifying framework for non-monotonic inference" by contributing to the development of "a general domain-independent criterion for inferring preferences from theories" [12]. The mathematical complexity of Geffner's definitions may reflect the generality of his goal. By comparison, in UCL both aim and means are modest. It appears that UCL can be embedded in Geffner's formalism, perhaps with some minor technical modifications, but we do not pursue this possibility here. The rewards would be minimal, given the differences in emphasis, and in mathematical machinery.

It appears that Geffner's proposal was inspired by Judea Pearl's investigation of the distinction between causal and non-causal grounds in general default reasoning and probabilistic reasoning, as developed in [45] and many subsequent publications. Consideration of the relationship of UCL to this body of work is beyond the scope of this note.

In recent years, many researchers have put forward proposals for causal theories of action and change [12,10,4,2,33,38,49,22,48,21,30,39,50,51,18,19,5,16,52]. In this note we consider the relationship of UCL to only a few of these proposals. As described previously, UCL can be understood as an extension of the causal theories approach of [39], where Norman McCain and the current author introduced so-called "causal laws" of the form $\phi \Rightarrow \psi$, with the intended reading "necessarily, if $\phi$ then the fact that $\psi$ is caused." Here we show that such causal laws can be translated in UCL as $\phi \supset \mathsf{C}\psi$, thus providing a more adequate semantic account of them. We go on to develop in some detail the close relationship between such causal laws and the circumscriptive approach to "causal laws" of Lin [33,34]. Along the same lines, we show that the "static causal laws" of [38,51] correspond to UCL formulas of the form $\mathsf{C}\phi \supset \mathsf{C}\psi$.

A useful fragment of UCL has a concise translation into classical propositional logic, via the "literal completion" method of [39], which is closely related to the well-known Clark completion method for logic programs [8]. A larger fragment of UCL corresponds closely to logic programming under the answer set semantics of Gelfond and Lifschitz [14]. These translations allow automated reasoning with UCL, using publicly available, fast improving satisfiability solvers and logic programming systems. Experimental results reported in [40] demonstrate that such an approach can be used to solve (what are currently) hard classical planning problems.

The contributions of this note can be summarized as follows. It introduces UCL, a mathematically simple modal nonmonotonic logic designed for representing commonsense knowledge about actions. By establishing relationships with previous proposals, it shows how a variety of causal theories of action can be expressed in UCL. By relating these proposals to a single logical framework, it contributes to the ongoing investigation of the relationships between various approaches. Finally, it relates UCL to some well-known nonmonotonic formalisms, and identifies methods for carrying out automated reasoning on the basis of UCL theories.

We proceed as follows. Section 2 defines propositional UCL, the fragment primarily investigated in this note. Section 3 presents preliminary examples. Section 4 shows that UCL extends the causal theories formalism of [39], and Section 5 describes the general method of formalizing action domains inherited from that paper. Section 6 relates UCL to default logic and logic programming, and Section 7 considers some results thereby inherited from [38,46,51]. Section 8 shows that a fragment of UCL can be nicely reduced to circumscriptive theories, and Section 9 explores the relationship between UCL and the circumscriptive action theories of Lin [33,34]. In Section 10, we consider the relationship of UCL to autoepistemic logic. In Section 11, we extend UCL to allow quantifiers. In Section 12, we show that UCL extends the nonpropositional causal theories of Lifschitz [30], which, in turn, extend the propositional causal theories discussed in Section 4. Section 13 consists of concluding remarks.

## 2 Propositional UCL

Begin with a set of propositional symbols (atoms)—the signature of our language. For convenience, we assume that the language includes zero-place logical connectives $\top$ and $\bot$ such that $\top$ and $\neg\bot$ are tautological. A *literal* is an atom or its negation. We identify each interpretation with the set of literals true in it. UCL *formulas* are defined as usual for a modal propositional language with single unary modal operator $\mathsf{C}$. A formula is *nonmodal* if $\mathsf{C}$ does not occur in it. A UCL *theory* is a set of UCL formulas.

The main semantic definition (of a "causally explained" interpretation) is obtained by imposing a fixpoint condition on S5 modal logic. Thus, a UCL *structure* is a pair $(I, S)$ such that $I$ is an interpretation, and $S$ is a set of interpretations to which $I$ belongs. The truth of a UCL sentence in a UCL structure is defined by the standard recursions over the propositional connectives, plus the following two conditions.

$$(I, S) \models p \quad \text{iff} \quad I \models p \quad \text{(for any atom } p\text{)}$$

$$(I, S) \models \mathsf{C}\phi \quad \text{iff} \quad \text{for all } I' \in S,\ (I', S) \models \phi$$

Given a UCL theory $T$, we write $(I, S) \models T$ to mean that $(I, S) \models \phi$, for every $\phi \in T$. In this case, we say that $(I, S)$ is a *model* of $T$. We also say that $(I, S)$ is an *$I$-model* of $T$, emphasizing the distinguished interpretation $I$.

**Main Definition.** Let $T$ be a UCL theory. An interpretation $I$ is *causally explained* by $T$ if $(I, \{I\})$ is the unique $I$-model of $T$.

We distinguish three entailment relations. The first two—classical propositional entailment and propositional S5 entailment—are standard, monotonic relations. The third—UCL entailment—is defined as follows. For any UCL theory $T$ and nonmodal formula $\phi$, we write $T \mathrel{|\!\!\approx} \phi$ to say that $\phi$ is true in every interpretation causally explained by $T$.

## 3 Examples

Let $T_1$ be the UCL theory with one formula

$$p \supset \mathsf{C}p$$

in the language with a single atom $p$. Let $I_1$ be the interpretation $\{p\}$. The structure $(I_1, \{I_1\})$ is the unique $I_1$-model of $T_1$, so $I_1$ is causally explained by $T$. The only other interpretation is $I_2 = \{\neg p\}$. Since $(I_2, \{I_1, I_2\}) \models T_1$, $I_2$ is not causally explained by $T_1$. Therefore, $T_1 \mathrel{|\!\!\approx} p$.

Notice that it is essential that the language of $T_1$ include only the atom $p$. If the language of $T_1$ were extended to include a second atom $q$, there would no longer be any causally explained interpretations, since, intuitively, $T_1$ includes no formula expressing conditions under which $q$ is caused. Notice that we can obtain a conservative extension of $T_1$ by adding, for instance, the formula $\mathsf{C}q \vee \mathsf{C}\neg q$. The resulting UCL theory has the same UCL-consequences in the language of $T_1$, since its causally explained interpretations are $\{p, q\}$ and $\{p, \neg q\}$. In fact, it also has the same S5-consequences in the language of $T_1$.

Let $T_2$ be the UCL theory obtained by adding to $T_1$ the formula

$$\neg p \supset \mathsf{C}\neg p\,.$$

Both $I_1$ and $I_2$ are causally explained by $T_2$. Therefore, $T_2 \not\approx p$, which illustrates the nonmonotonicity of UCL.

Let $T_3$ be the UCL theory obtained from $T_2$ by adding the atom $q$ to the language, and also adding the formula

$$\mathsf{C}(q \equiv p)\,.$$

The interpretations $\{p, q\}$ and $\{\neg p, \neg q\}$ are both causally explained by $T_3$. No others are.

This last example illustrates the following general phenomenon. We obtain a definitional extension $T'$ of a UCL theory $T$ by adding a new atom $p$ to the signature and also adding an *explicit definition* of $p$—a formula of the form

$$\mathsf{C}\,(p \equiv \phi) \tag{4}$$

where $\phi$ is a nonmodal formula in which $p$ does not occur. It is not difficult to verify that $T'$ is a conservative extension of $T$. Moreover, one can replace any formula equivalent to $\phi$ by $p$ anywhere in $T'$, except in (4), without affecting the models of $T'$, or, therefore, the causally explained interpretations.

## 4 Causal Theories in UCL

In [39], Norman McCain and the current author defined the so-called causal theories formalism, together with a general method of formalizing action domains as causal theories. This section recalls the definition of a causal theory, and specifies a simple translation from causal theories into UCL. It also describes the "literal completion" method inherited by UCL from causal theories, which provides a concise translation of a fragment of UCL into classical propositional logic.

### *4.1 McCain and Turner's Causal Theories*

A *causal law* is an expression of the form

$$\phi \Rightarrow \psi \tag{5}$$

where $\phi$ and $\psi$ are (nonmodal) formulas. By a *causal theory* we mean a set of causal laws. We emphasize that (5) is not the material conditional $\phi \supset \psi$. In fact, we will show that $\phi \Rightarrow \psi$ can be translated in UCL as

$$\phi \supset \mathsf{C}\psi\,. \tag{6}$$

As in UCL, the main definition in the language of causal theories is that of a causally explained interpretation. Let $D$ be a causal theory and $I$ an interpretation. Define

$$D^I = \{\, \psi \,:\, \text{for some } \phi,\ \phi \Rightarrow \psi \in D \text{ and } I \models \phi \,\}\,.$$

We say that $I$ is *causally explained* by $D$ if $I$ is the unique model of $D^I$.

### *4.2 Embedding Causal Theories in UCL*

Given a causal theory $D$, let $ucl(D)$ denote the UCL theory obtained from $D$ by replacing each causal law (5) with the corresponding UCL formula (6).

**Theorem 1** *For any causal theory $D$, an interpretation $I$ is causally explained by $D$ if and only if $I$ is causally explained by $ucl(D)$.*

**Lemma 2** *For any causal theory $D$ and UCL structure $(I,S)$, $(I,S) \models ucl(D)$ if and only if, for all $I' \in S$, $I' \models D^I$.*

**Proof.** The lemma follows easily from the following observation. For every $\phi \Rightarrow \psi \in D$, the following two conditions are equivalent.

- $(I,S) \models \phi \supset \mathsf{C}\psi$
- If $I \models \phi$, then, for all $I' \in S$, $I' \models \psi$.

**Proof of Theorem 1.** ($\Longrightarrow$) Assume that $I$ is the unique model of $D^I$. By Lemma 2, $(I,\{I\}) \models ucl(D)$. Let $S$ be a superset of $\{I\}$ s.t. $(I,S) \models ucl(D)$. By Lemma 2, for all $I' \in S$, $I' \models D^I$. It follows that $S = \{I\}$, so $(I,\{I\})$ is the unique $I$-model of $ucl(D)$.

($\Longleftarrow$) Assume that $(I,\{I\})$ is the unique $I$-model of $ucl(D)$. By Lemma 2, $I \models D^I$. Assume that $I' \models D^I$. By Lemma 2, $(I,\{I,I'\}) \models ucl(D)$. It follows that $I = I'$, so $I$ is the unique model of $D^I$.

### *4.3 UCL in Classical Propositional Logic*

Here we define the class of "definite" UCL theories, which inherit from ("definite") causal theories a concise translation into classical propositional logic, by the literal completion method [39].

A UCL formula is *definite* if it is nonmodal or has the form

$$\phi \supset \mathsf{C}L \tag{7}$$

where $\phi$ is nonmodal and $L$ is a literal. A UCL theory $T$ is *definite* if

- each of its formulas is definite, and
- for every literal $L$, $T$ has finitely many formulas (7) with consequent $\mathsf{C}L$.

Let $T$ be a definite UCL theory. By the *literal completion* of $T$ we mean the classical propositional theory obtained by an elaboration of the Clark completion method [8], as follows. For each literal $L$ in the language of $T$, include the formula

$$L \equiv (\phi_1 \vee \cdots \vee \phi_n) \tag{8}$$

where $\phi_1, \ldots, \phi_n$ are the antecedents of the formulas of form (7) in $T$ with consequent $\mathsf{C}L$. (Of course, if no formula (7) in $T$ has consequent $\mathsf{C}L$, then formula (8) becomes $L \equiv \bot$.) Include also all nonmodal formulas from $T$.

**Theorem 3** *An intepretation is causally explained by a definite UCL theory $T$ if and only if it is a model of the literal completion of $T$.*

This result follows immediately, by Theorem 1, from Proposition 1 of [40]. In that paper, classical propositional theories obtained by literal completion from causal theories are used to solve hard classical planning problems from [27], demonstrating the potential effectiveness of this approach to automated reasoning with UCL.

## 5 Causal Theories of Action in UCL

By Theorem 1, UCL inherits the general approach to action formalization from [39]. Here we review it, in the UCL setting.

### *5.1* $\mathcal{L}(\mathbf{F}, \mathbf{A}, \mathbf{T})$ *Languages*

It is convenient to specify the underlying propositional signature by means of three pairwise-disjoint sets: a nonempty set $\mathbf{F}$ of *fluent names*, a set $\mathbf{A}$ of *action names*, and a nonempty set $\mathbf{T}$ of *time names*, corresponding to the natural numbers or an initial segment of them. The atoms of the language $\mathcal{L}(\mathbf{F}, \mathbf{A}, \mathbf{T})$ are divided into two classes, defined as follows. The *fluent atoms* are expressions of the form $f_t$ such that $f \in \mathbf{F}$ and $t \in \mathbf{T}$. Intuitively, $f_t$ is true if and only if the fluent $f$ holds at time $t$. The *action atoms* are expressions of the form $a_t$ such that $a \in \mathbf{A}$ and $t + 1 \in \mathbf{T}$. Intuitively, $a_t$ is true if and only if the action $a$ occurs at time $t$. A *fluent literal* is a fluent atom or its negation. A *fluent formula* is a propositional combination of fluent atoms.

### *5.2* $\mathcal{L}(\mathbf{F}, \mathbf{A}, \mathbf{T})$ *Domain Descriptions*

We illustrate the approach by formalizing a slight elaboration of Lin's Suitcase domain [33] in which there is a suitcase with two latches, each of which may be in either of two positions, up or down. The suitcase is spring-loaded so that whenever both latches are in the up position the suitcase is caused to be open. We model the opening of the suitcase as a static effect (as Lin does); that is, we do not model a state of the domain in which both latches are up but the suitcase is not (yet) open.

We take time names corresponding to the natural numbers, and we choose fluent names and action names as follows.

$$\text{Fluents} \begin{cases} \mathit{Up}(L_1) : \text{the first latch is up} \\ \mathit{Up}(L_2) : \text{the second latch is up} \\ \mathit{IsOpen} : \text{the suitcase is open} \end{cases}$$

$$\text{Actions} \begin{cases} \mathit{Toggle}(L_1) : \text{toggle the first latch} \\ \mathit{Toggle}(L_2) : \text{toggle the second latch} \\ \mathit{Close} : \text{close the suitcase} \end{cases}$$

Given our choice of language, the Suitcase domain can be partially formalized by the following four schemas, where $l$ is a metavariable ranging over $\{L_1, L_2\}$.

$$\mathit{Toggle}(l)_t \wedge \mathit{Up}(l)_t \supset \mathsf{C}\neg \mathit{Up}(l)_{t+1} \tag{9}$$
$$\mathit{Toggle}(l)_t \wedge \neg \mathit{Up}(l)_t \supset \mathsf{C}\, \mathit{Up}(l)_{t+1} \tag{10}$$
$$\mathit{Close}_t \supset \mathsf{C}\neg \mathit{IsOpen}_{t+1} \tag{11}$$

$$Up(L_1)_t \wedge Up(L_2)_t \supset \mathsf{C}IsOpen_t \tag{12}$$

According to schemas (9) and (10), whenever a latch is toggled at a time $t$ it is caused to be in the opposite state at time $t+1$. Schema (11) says that whenever the suitcase is closed at a time $t$ it is caused to be not open at $t+1$. Schema (12) says that whenever both latches are up at a time $t$ the suitcase is caused to be open also at $t$. Schemas (9)–(11) express "dynamic causal laws." Schema (12) expresses a "static causal law."

The UCL theory (9)–(12) is an incomplete description of the Suitcase domain because it does not represent sufficient conditions for certain facts being caused: namely, facts preserved by inertia, facts about the initial situation, and facts about which actions occur (and when). The following schemas provide a standard way to complete the description.

In the following two schemas, $a$ is a metavariable for action names.

$$a_t \supset \mathsf{C}a_t \tag{13}$$
$$\neg a_t \supset \mathsf{C}\neg a_t \tag{14}$$

Schema (13) says that the occurrence of an action $a$ at a time $t$ is caused whenever $a$ occurs at $t$. Schema (14) says that the non-occurrence of an action $a$ at a time $t$ is caused whenever $a$ does not occur at $t$. In effect, by these schemas we represent that facts about action occurrences are exogenous to the theory.

In the following two schemas, $f$ is a metavariable for fluent names.

$$f_0 \supset \mathsf{C}f_0 \tag{15}$$
$$\neg f_0 \supset \mathsf{C}\neg f_0 \tag{16}$$

In effect, by these schemas we represent that facts about the initial values of fluents may be exogenous to the theory.

By a *fluent designating formula* we mean a propositional combination of fluent names. Given a fluent designating formula $\sigma$ and a time name $t$, we write $\sigma_t$ to stand for the fluent formula obtained from $\sigma$ by simultaneously replacing each occurrence of each fluent name $f$ by the fluent atom $f_t$.

Let $\mathbf{I}$ be a set of fluent designating formulas. We express that the fluents designated by the formulas in $\mathbf{I}$ are inertial by writing the following schema, where $\sigma$ is a metavariable ranging over $\mathbf{I}$.

$$\sigma_t \wedge \sigma_{t+1} \supset \mathsf{C}\sigma_{t+1} \tag{17}$$

According to schema (17), whenever a fluent designated in $\mathbf{I}$ holds at two successive times, its truth at the second time is taken to be caused simply by virtue of its persistence. For the Suitcase domain, we take **I** to be the set of all fluent names and their negations. [1]

Schemas (9)–(17) express the complete UCL theory $T_4$ for the Suitcase domain. Schemas (9)–(12) are domain specific. We call the remaining schemas (13)–(17) *standard schemas*. Intuitively, the standard schemas exempt specific classes of facts from the principle of universal causation.

Let $I$ be the interpretation characterized below.

| | | | |
|---|---|---|---|
| • $\neg Toggle(L_1)_0$ | • $\neg Toggle(L_1)_1$ | • $\neg Toggle(L_1)_2$ | $\cdots$ |
| • $Toggle(L_2)_0$ | • $\neg Toggle(L_2)_1$ | • $\neg Toggle(L_2)_2$ | $\cdots$ |
| • $\neg Close_0$ | • $\neg Close_1$ | • $\neg Close_2$ | $\cdots$ |
| • $Up(L_1)_0$ | • $Up(L_1)_1$ | • $Up(L_1)_2$ | $\cdots$ |
| • $\neg Up(L_2)_0$ | $Up(L_2)_1$ | • $Up(L_2)_2$ | $\cdots$ |
| • $\neg IsOpen_0$ | $IsOpen_1$ | • $IsOpen_2$ | $\cdots$ |

Interpretation $I$ specifies, for all actions $a$ and times $t$, whether or not $a$ occurs at $t$, and, for all fluents $f$ and times $t$, whether or not $f$ holds at $t$. Here, exactly one action occurs—the toggling of the second latch at time 0—and, intuitively, it results in the suitcase being open at time 1. (The ellipses indicate that after time 2 no action occurs and no fluent changes its value. The bullets indicate literals that are "explained" by the standard schemas.) It is not difficult to see that $I$ is causally explained by $T_4$.

The following formula is a UCL-consequence of $T_4$.

$$Up(L_1)_0 \wedge Up(L_2)_0 \wedge Close_0 \supset Toggle(L_1)_0 \vee Toggle(L_2)_0$$

In general, when both latches are up, it is impossible to perform *only* the action of closing the suitcase; one must also concurrently toggle at least one of the latches. If this seems unintuitive, recall that we have chosen to model the suitcase being open as a static effect of the latches being up, so there is no time in any causally possible world at which both latches are up and the suitcase is closed.

[1] Thus, the inertia laws for the Suitcase domain can also be represented by the pair of schemas

$$f_t \wedge f_{t+1} \supset \mathsf{C} f_{t+1}$$
$$\neg f_t \wedge \neg f_{t+1} \supset \mathsf{C} \neg f_{t+1}$$

where $f$ is a metavariable for fluent names. In other cases, there may be inertial fluents that are not designated by fluent names or their negations, and, conversely, there may be fluent names or negations of fluent names that do not designate inertial fluents. We will see an example of this in Section 5.3.3. A still more general form of inertia is discussed in Section 5.3.5.

### 5.3 Additional Expressive Possibilities

The previous example demonstrates that UCL can be used to represent some standard features of action domains, such as indirect effects of actions, implied action preconditions and concurrent actions. Next we briefly describe a few of the additional expressive possibilities of the approach.

#### 5.3.1 Ramification and Qualification Constraints

Ramification and qualification constraints, in the sense of Lin and Reiter [35], are formalized by schemas of the forms $\mathsf{C}\sigma_t$ and $\sigma_t$ respectively, where $\sigma$ (the "state constraint") is a fluent designating formula. We will consider this claim in more detail in Section 7.

#### 5.3.2 Nondeterministic Actions

The semantics of UCL rests on the principle of universal causation, according to which every fact is caused. Intuitively, in the case of a nondeterministic action, there is no cause for one of its possible effects rather than another. We have already seen, however—in standard schemas (13) through (17)—that there are ways of effectively exempting facts from the principle of universal causation. We can use laws of a similar form to describe nondeterministic actions. For instance, coin tossing can be described (in part) as follows.

$$Toss_t \wedge Heads_{t+1} \supset \mathsf{C}Heads_{t+1} \tag{18}$$
$$Toss_t \wedge \neg Heads_{t+1} \supset \mathsf{C}\neg Heads_{t+1} \tag{19}$$

Intuitively, according to schemas (18) and (19), for every time $t$, $Toss_t$ renders $Heads_{t+1}$ exogenous. We'll consider some related results in Section 9.3.

#### 5.3.3 Defined Fluents

Given an $\mathcal{L}(\mathbf{F}, \mathbf{A}, \mathbf{T})$ domain description, we add a defined fluent $f$ ($f \notin \mathbf{F}$) by first adding $f$ to the set of fluent names and then defining $f$ by means of a schema

$$\mathsf{C}(f_t \equiv \sigma_t)$$

where $\sigma$ is a fluent designating formula that doesn't mention $f$. It is important that the set $\mathbf{I}$ used to designate the inertial fluents is not altered in this process. Intuitively speaking, the defined fluent inherits any inertial properties it may have from its definiens. The correctness of this method of introducing defined fluents follows from the remarks on definitional extension in Section 3.

### 5.3.4 *Delayed Effects and Things that Change by Themselves*

Because we refer explicitly to time points in our action descriptions, we may, if we wish, describe actions with delayed effects. We may also model things that change by themselves. This we can do simply by writing causal laws that relate fluents at different times, without mentioning any actions. (Alternatively, we may explicitly introduce "events," which, like actions, can be conceived to be the causes of change.)

Consider an example from [40] along these lines, involving the dynamic mechanism of falling dominos. We wish to describe the chain reaction of four dominos falling over one after the other, after the first domino is tipped over. We describe the direct effect and action precondition of the *Tip* action by writing

$$Tip_t \supset \mathsf{C}\neg Up(1)_{t+1} \tag{20}$$

$$Tip_t \supset Up(1)_t\,. \tag{21}$$

According to (20), *Tip* is the action of tipping over the first domino. According to (21), it can only be done if the first domino is standing upright. We describe the chain reaction mechanism as follows, where $d$ is a metavariable ranging over the numbers $1, 2, 3$.

$$Up(d)_t \wedge \neg Up(d)_{t+1} \supset \mathsf{C}\neg Up(d+1)_{t+2} \tag{22}$$

Notice that (22) does not mention an action. It describes dynamic change involving three distinct time points. Roughly speaking, if domino $d$ falls in the interval from $t$ to $t+1$, then domino $d+1$ is caused to fall in the interval from $t+1$ to $t+2$.

### 5.3.5 *Generalized Commonsense Law of Inertia*

The fact that the commonsense law of inertia can be expressed straightforwardly in UCL makes it easy to generalize, as follows. Rather than supposing that things tend to stay the same, we can imagine more generally that they tend to change in particular ways. That is, there is a course that nature would follow, in the absence of interventions.

As an example, consider the Pendulum domain from [20]. In the course of nature, a pendulum swings back and forth, from right to left and back again. However, at any time an agent can intervene by holding the pendulum in its current location. When the agent no longer holds it, the pendulum resumes its natural course, swinging back and forth. The effects of the action *Hold* are specified by writing

$$Hold_t \wedge Right_t \supset \mathsf{C}Right_{t+1}$$

$$Hold_t \wedge \neg Right_t \supset \mathsf{C}\neg Right_{t+1}\,.$$

The behavior of the pendulum in the absence of interventions is described by writing

$$\neg Right_t \wedge Right_{t+1} \supset \mathsf{C} Right_{t+1} \tag{23}$$

$$Right_t \wedge \neg Right_{t+1} \supset \mathsf{C} \neg Right_{t+1}. \tag{24}$$

Like the standard inertia schema (17), schemas (23) and (24) describe a course of nature. Here the course of nature is dynamic rather than static, but otherwise there are clear similarities. Like (17), axioms (23) and (24) allow for the possibility that the course of nature may be overridden by the effects of actions, and they do so without mentioning facts about the non-occurrence of actions as preconditions. So, in essence, these axioms solve the frame problem for the dynamic fluent *Right* in the same way that standard inertia axioms solve the frame problem for inertial fluents.

## 6 UCL and Default Logic

In this section, we establish the close mathematical relationship between UCL and default logic [47]. More precisely, we consider a generalization of default logic, called disjunctive default logic [17], which includes Reiter's default logic as a special case. The semantics of a disjunctive default theory is given in terms of its extensions, which are logically closed sets of (nonmodal) formulas that satisfy a certain fixpoint condition. Although an extension may be inconsistent, or incomplete (that is, there may be an atom $p$ such that neither $p$ nor $\neg p$ belong to it), we will be interested in the special case of extensions that are both consistent and complete, since it is these extensions that correspond to interpretations.

We will specify a translation from disjunctive default logic to UCL such that the complete, consistent extensions correspond to the causally explained interpretations. The translation is invertible, so there is a strong sense in which UCL is equivalent to disjunctive default logic, restricted to the special case of complete, consistent extensions.

Since disjunctive logic programming under the answer set semantics [14] can be understood as a special case of disjunctive default logic, as established in [17], this translation also yields a correspondence between UCL and logic programming, which makes possible automated reasoning using fast improving, publicly available systems such as SMODELS [44], DLV [9] and DeReS [7]. (In fact, DeReS can even work directly with default theories.)

### *6.1 Disjunctive Default Logic*

Here we recall definitions from [17].

A *disjunctive default rule* is an expression of the form

$$\frac{\alpha : \beta_1, \ldots, \beta_m}{\gamma_1 | \cdots | \gamma_n} \tag{25}$$

where all of $\alpha, \beta_1, \ldots, \beta_m, \gamma_1, \ldots, \gamma_n$ are (nonmodal) formulas ($m \geq 0, n \geq 1$).

A *disjunctive default theory* is a set of disjunctive default rules. Let $D$ be a disjunctive default theory and $E$ a set of formulas. Define $D^E$ as follows.

$$D^E = \left\{ \frac{\alpha}{\gamma_1 | \cdots | \gamma_n} : \frac{\alpha : \beta_1, \ldots, \beta_m}{\gamma_1 | \cdots | \gamma_n} \in D \text{ and } \neg\beta_1, \ldots, \neg\beta_m \notin E \right\}$$

A set $E'$ of formulas is *closed under* $D^E$ if, for every member of $D^E$, if $\alpha \in E'$ then at least one of $\gamma_1, \ldots, \gamma_n$ belongs to $E'$. We say $E$ is an *extension* for $D$ if $E$ is minimal among sets closed under propositional logic and closed under $D^E$. We say $E$ is *complete* if, for every atom $p$, either $p \in E$ or $\neg p \in E$. Notice that, for the purpose of computing complete, consistent extensions, the *justifications* $\beta_1, \ldots, \beta_m$ of a disjunctive default rule can be safely replaced with their conjunction.

Reiter's default logic corresponds to the special case when $n = 1$.[2]

### *6.2 UCL and Disjunctive Default Logic*

Given a disjunctive default theory $D$, let *ucl*$(D)$ be the UCL theory obtained from $D$ by replacing each disjunctive default rule (25) with the UCL formula

$$\mathsf{C}\alpha \wedge \beta_1 \wedge \cdots \wedge \beta_m \supset \mathsf{C}\gamma_1 \vee \cdots \vee \mathsf{C}\gamma_n \, . \tag{26}$$

It is a fact of propositional S5 modal logic that every theory is equivalent to one in which every formula has the form (26).[3] Thus, every UCL theory is equivalent to one that can be obtained by this translation from disjunctive default logic.

[2] In Reiter's formulation, a default theory is a pair $(D, W)$, where the second component $W$ is a set of formulas. Here we suppress the second component, since every $\phi \in W$ can be equivalently represented by the justification-free rule $\frac{\top :}{\phi}$.

[3] This follows, for instance, from the MCNF Theorem in [24].

Given a set $S$ of interpretations, let $Th(S)$ denote the set of nonmodal formulas true in all members of $S$. Given a set $\Gamma$ of nonmodal formulas, let $Mod(\Gamma)$ denote the set of interpretations that satisfy all members of $\Gamma$.

**Theorem 4** *For any disjunctive default theory $D$ and any interpretation $I$, $Th(\{I\})$ is an extension for $D$ if and only if $I$ is causally explained by $ucl(D)$.*

**Lemma 5** *For any disjunctive default theory $D$ and UCL structure $(I, S)$, $(I, S) \models ucl(D)$ if and only if $Th(S)$ is closed under $D^{Th(\{I\})}$.*

**Proof.** ($\Longrightarrow$) Assume that $(I, S) \models ucl(D)$. Consider any rule $\frac{\alpha}{\gamma_1 | \cdots | \gamma_n}$ in $D^{Th(\{I\})}$ such that $\alpha \in Th(S)$. We must show that at least one of $\gamma_1, \ldots, \gamma_n$ is in $Th(S)$. We know there is a rule $\frac{\alpha : \beta_1, \ldots, \beta_m}{\gamma_1 | \cdots | \gamma_n}$ in $D$ such that $I$ satisfies all of $\beta_1, \ldots, \beta_m$. It follows that $\mathsf{C}\alpha \wedge \beta_1 \wedge \cdots \wedge \beta_m \supset \mathsf{C}\gamma_1 \vee \cdots \vee \mathsf{C}\gamma_n$ is in $ucl(D)$, and that $(I, S) \models \beta_1 \wedge \cdots \wedge \beta_m$. Since $\alpha \in Th(S)$, $(I, S) \models \mathsf{C}\alpha$. Since $(I, S) \models ucl(D)$, we can conclude that $(I, S) \models \mathsf{C}\gamma_1 \vee \cdots \vee \mathsf{C}\gamma_n$. Thus, there is an $i \in \{1, \ldots, n\}$ such that $(I, S) \models \mathsf{C}\gamma_i$, and consequently $\gamma_i \in Th(S)$.

($\Longleftarrow$) Assume that $Th(S)$ is closed under $D^{Th(\{I\})}$. Consider any formula $\mathsf{C}\alpha \wedge \beta_1 \wedge \cdots \wedge \beta_m \supset \mathsf{C}\gamma_1 \vee \cdots \vee \mathsf{C}\gamma_n$ from $ucl(D)$ whose antecedent is satisfied by $(I, S)$. We must show that $(I, S)$ satisfies at least one of $\mathsf{C}\gamma_1, \ldots, \mathsf{C}\gamma_n$. We know that $\frac{\alpha : \beta_1, \ldots, \beta_m}{\gamma_1 | \cdots | \gamma_n}$ belongs to $D$. Because $(I, S) \models \beta_1 \wedge \cdots \wedge \beta_m$, we know that $\neg\beta_1, \ldots, \neg\beta_m \notin Th(\{I\})$. Therefore, $\frac{\alpha}{\gamma_1 | \cdots | \gamma_n}$ belongs to $D^{Th(\{I\})}$. And because $(I, S) \models \mathsf{C}\alpha$, we know that $\alpha \in Th(S)$. Since $Th(S)$ is closed under $D^{Th(\{I\})}$, we can conclude that at least one of $\gamma_1, \ldots, \gamma_n$ is in $Th(S)$. It follows that $(I, S)$ satisfies at least one of $\mathsf{C}\gamma_1, \ldots, \mathsf{C}\gamma_n$.

**Proof of Theorem 4.** ($\Longrightarrow$) Assume that $Th(\{I\})$ is an extension for $D$. We know by Lemma 5 that $(I, \{I\}) \models ucl(D)$. Let $S$ be a superset of $\{I\}$ such that $(I, S) \models ucl(D)$. By Lemma 5, $Th(S)$ is closed under $D^{Th(\{I\})}$. Because $I \in S$, $Th(S) \subseteq Th(\{I\})$. And since $Th(\{I\})$ is a minimal among sets closed under $D^{Th(\{I\})}$, we can conclude that $Th(S) = Th(\{I\})$, from which it follows that $S = \{I\}$. So $(I, \{I\})$ is the unique $I$-model of $ucl(D)$. That is, $I$ is causally explained by $ucl(D)$.

($\Longleftarrow$) Assume $I$ is causally explained by $ucl(D)$. So $(I, \{I\}) \models ucl(D)$. By Lemma 5, $Th(\{I\})$ is closed under $D^{Th(\{I\})}$. Let $E$ be a subset of $Th(\{I\})$ that is closed under propositional logic and under $D^{Th(\{I\})}$. By Lemma 5, $(I, Mod(E)) \models ucl(D)$. Since $(I, \{I\})$ is the unique $I$-model of $ucl(D)$, we have $Mod(E) = \{I\}$. It follows that $E = Th(\{I\})$. We can conclude that $Th(\{I\})$ is a minimal set closed under propositional logic and under $D^{Th(\{I\})}$. That is, $Th(\{I\})$ is an extension for $D$.

In the statement of Theorem 4, we restrict attention to extensions that can be expressed in the form $Th(\{I\})$, where $I$ is an interpretation. That is, we consider only complete, consistent extensions. This restriction can of course be expressed in the default theory itself, simply by adding the justification $\top$ to each justification-free default rule (to guarantee that all extensions are consistent), and also adding the default rule

$$\frac{:p, \neg p}{\bot}$$

for each atom $p$ in the language (to guarantee that all extensions are complete).

## 7 More on Actions and Change in UCL

On the basis of Theorem 4, UCL inherits results on reasoning about action and theory update from [38,46,51]. Here we describe some of these.

In [38], we gave a uniform account of ramification and qualification constraints (in the sense of Lin and Reiter [35]), using inference rules to express static causal laws. There the central definition identifies the states that can result from performing an action with "direct" effect $E$ in a state $S$, in a world with causal relationships characterized by a set $C$ of inference rules. This definition of "possible next states" is given by a fixpoint condition that can be seen to reflect the principle of universal causation, understood in a less general fashion. Here we recall the definition from [38], beginning with a few preliminary definitions.

Consider a propositional language in which the atoms are fluent names. Intuitively, an interpretation of this language represents a state of the world. We will understand an inference rule $\phi/\psi$ in this language to express the static causal law "$\psi$ is caused whenever $\phi$ is." Let $\Gamma$ be a set of formulas. We say that $\Gamma$ is *closed under* a set $C$ of inference rules, if, for every inference rule $\phi/\psi$ in $C$, if $\phi$ belongs to $\Gamma$, so does $\psi$. We write $\Gamma \vdash_C \phi$ if $\phi$ belongs to the least set of formulas closed under $C$ and closed under propositional logic.

Let $C$ be a set of inference rules. Let $S$ and $S'$ be interpretations. (Recall that we identify an interpretation with the set of literals true in it.) Let $E$ be a set of formulas. We say that *$S'$ can result from causing $E$ in $S$ relative to $C$* if

$$S' = \{\, L \,:\, (S \cap S') \cup E \vdash_C L \,\}$$

where $L$ stands exclusively for literals.

In [46], this definition was explored in the more abstract setting of theory update. There we showed (Theorem 8.1) that it can be simply embedded in default logic, using inertia rules closely related to the UCL inertia axioms from Section 5, as follows.

**Proposition 6** *Let $C$ be a set of inference rules, $S$ an interpretation, and $E$ a set of formulas. An interpretation $S'$ can result from causing $E$ in $S$ relative to $C$ if and only if $Th(S')$ is an extension of the default theory*

$$\left\{\frac{\top : L}{L} : L \in S\right\} \cup \left\{\frac{\top :}{\phi} : \phi \in E\right\} \cup \left\{\frac{\phi :}{\psi} : \phi/\psi \in C\right\}.$$

Using this result, along with Theorem 4, we immediately obtain the following.

**Corollary 7** *Let $C$ be a set of inference rules, $S$ an interpretation, and $E$ a set of formulas. An interpretation $S'$ can result from causing $E$ in $S$ relative to $C$ if and only if $S'$ is causally explained by the UCL theory*

$$\{L \supset \mathsf{C}L : L \in S\} \cup \{\mathsf{C}\phi : \phi \in E\} \cup \{\mathsf{C}\phi \supset \mathsf{C}\psi : \phi/\psi \in C\}. \tag{27}$$

Here the inertial property of each literal $L$ that is true in the initial state $S$ is captured by the formula $L \supset \mathsf{C}L$, which can be understood to say that $L$ is caused in a subsequent state $S'$ if it is true in $S'$.

Corollary 7 shows that the static causal laws from [38] correspond to UCL formulas of the form

$$\mathsf{C}\phi \supset \mathsf{C}\psi$$

where $\phi$ and $\psi$ are nonmodal formulas.

Similarly, UCL formulas of the form

$$\mathsf{C}\phi$$

where $\phi$ is nonmodal, correspond to traditional ramification constraints. Here we validate this claim with respect to Winslett's classic, minimal-change definition of theory update [53], which can be stated as follows.

Let $\Gamma$ be a set of formulas and $S$ an interpretation. We say a model $S'$ of $\Gamma$ is a *minimal-change update* of $S$ by $\Gamma$ if there is no model $S''$ of $\Gamma$ such that $S' \cap S$ is a proper subset of $S'' \cap S$.

**Proposition 8** *Let $\Gamma$ be a set of formulas and $S$ an interpretation. An interpretation $S'$ is a minimal-change update of $S$ by $\Gamma$ if and only if $S'$ is causally explained by the UCL theory*

$$\{L \supset \mathsf{C}L : L \in S\} \cup \{\mathsf{C}\phi : \phi \in \Gamma\}. \tag{28}$$

Proposition 8 follows immediately from Propositions 2 and 3 in [38], along with Corollary 7 above.

In a similar vein, it is clear that a nonmodal formula $\phi$, if added to (28) or (27), would act as a qualification constraint (again in the sense of Lin and Reiter [35]), simply eliminating the causally explained interpretations that fail to satisfy the constraint $\phi$.

In [51], a high-level action language $\mathcal{AC}$ was defined, in the manner of the action language $\mathcal{AR}_0$ of Kartha and Lifschitz [25], and an embedding into default logic was established. The language $\mathcal{AC}$ uses a slight extension of the definition of possible next states from [38], allowing for explicitly defined fluents and nondeterministic direct effects of actions. Like its predecessors $\mathcal{AR}_0$ and $\mathcal{A}$ [15], $\mathcal{AC}$ is based on the situation calculus [42]. The embedding of $\mathcal{AC}$ into default logic, along with Theorem 4 embedding default logic into UCL, shows that UCL can be used to express theories in the situation calculus, although we do not explore that possibility in this paper.

When actions are not always executable, situation calculus theories must include an additional predicate—typically named *Poss*—that characterizes when it is possible to perform an action in a situation, and when a situation can be reached by a sequence of actions. In general, when the preconditions for executability of actions are not explicitly given, the mathematics of *Poss* can be rather complex, as hinted at in footnote 5 of [33]. Even when action preconditions are given explicitly, as is required in the translation of $\mathcal{AC}$ into default logic in [51], the additional predicate and associated axioms complicate the theory. By contrast, when time has a linear structure and action occurrences are represented by propositions, as in the $\mathcal{L}(\mathbf{F}, \mathbf{A}, \mathbf{T})$ languages from Section 5.1, the question of whether an action can be performed in a particular situation need not be explicitly addressed in the theory. Instead, if an action can be performed in a situation, there will be, roughly speaking, some causally possible world in which it actually is. [4] For this reason, we find it convenient to suppress the role of the situation calculus when we describe the action theories from [51]. We will do the same when we consider in Section 9 the causal action theories of Lin [33,34].

In [51], inertia is expressed by default rules that, in light of Theorem 4, correspond (essentially) to the UCL schemas

$$\mathsf{C}f_t \wedge f_{t+1} \supset \mathsf{C}f_{t+1} \tag{29}$$

$$\mathsf{C}\neg f_t \wedge \neg f_{t+1} \supset \mathsf{C}\neg f_{t+1} \tag{30}$$

where $f$ is a metavariable ranging over fluent names. Notice that these inertia

[4] A similar idea is made precise in [40], where the executablity of a plan is defined for $\mathcal{L}(\mathbf{F}, \mathbf{A}, \mathbf{T})$ domain descriptions (expressed as causal theories).

schemas differ from the corresponding inertia schema (17) in Section 5.2. There $\mathsf{C}$ is not applied to $f_t$ and $\neg f_t$. The negative occurrences of $\mathsf{C}$ in (29) and (30) make them weaker (in S5) than their counterparts from Section 5.2. Similarly, static causal laws in [51] correspond to formulas of the form

$$\mathsf{C}\sigma_t \supset \mathsf{C}\sigma'_t \tag{31}$$

where $\sigma$ and $\sigma'$ are fluent designating formulas. Here again we see a negative occurrence of $\mathsf{C}$. Notice that formula (31) is an S5-consequence of $\sigma_t \supset \mathsf{C}\sigma'_t$, whose form corresponds to that of the static causal laws in Section 5.2.

The formalizations of action domains as default theories in [51], along with Theorem 4, suggests the following alternative UCL formalization of Lin's Suitcase domain. We begin with four domain specific schemas, where $l$ is metavariable ranging over $\{L_1, L_2\}$.

$$\mathsf{C}(\mathit{Toggle}(l)_t \wedge \mathit{Up}(l)_t) \supset \mathsf{C}\neg \mathit{Up}(l)_{t+1} \tag{32}$$
$$\mathsf{C}(\mathit{Toggle}(l)_t \wedge \neg \mathit{Up}(l)_t) \supset \mathsf{C}\,\mathit{Up}(l)_{t+1} \tag{33}$$
$$\mathsf{C}\,\mathit{Close}_t \supset \mathsf{C}\neg \mathit{IsOpen}_{t+1} \tag{34}$$
$$\mathsf{C}(\mathit{Up}(L_1)_t \wedge \mathit{Up}(L_2)_t) \supset \mathsf{C}\mathit{IsOpen}_t \tag{35}$$

To this we add the formulas given by standard schemas (13)–(16) as before, along with the alternative inertia schemas (29) and (30).

Notice that (32)–(35) can be obtained from the corresponding schemas in the Suitcase description from Section 5.2 simply by applying the modal operator $\mathsf{C}$ to each antecendent. Despite the new, negative occurrences of $\mathsf{C}$, it is possible to show that this formalization of the Suitcase domain yields the same causally explained interpretations as the UCL formalization considered in Section 5.2.

The following result establishes the inclusion in one direction. (We do not establish here the other direction of inclusion.)

**Proposition 9** *Let $T$ be a UCL theory in which $\mathsf{C}$ is applied only to nonmodal formulas. Assume also that all occurrences of the connective $\equiv$ have been eliminated from $T$. Let $T'$ be obtained from $T$ by replacing, in every formula of $T$, every negative occurrence of a subformula of the form $\mathsf{C}\phi$ with $\phi$. Any interpretation causally explained by $T$ is also causally explained by $T'$.*

This result can be proved on the basis of the following observations. For any UCL structure $(I, S)$ and nonmodal formula $\phi$, if $(I, \{I\}) \models \neg\mathsf{C}\phi$ then $(I, \{I\}) \models \neg\phi$, and if $(I, S) \not\models \neg\mathsf{C}\phi$ then $(I, S) \not\models \neg\phi$.

To see that, in general, the converse of Proposition 9 does not hold, consider UCL theories $T = \{\mathsf{C}p \supset \mathsf{C}p\}$ and $T' = \{p \supset \mathsf{C}p\}$, with $p$ the only atom in the language. The interpretation $\{p\}$ is causally explained by $T'$, but not by $T$.

## 8 A Subset of UCL in Circumscription

Let $T$ be a finite UCL theory, with a finite signature, in which the operator $\mathsf{C}$ is applied only to literals. In this section, we show that $T$ can be reduced to a circumscriptive theory $ct(T)$.[5]

The language $\mathcal{L}$ of $ct(T)$ is a second-order language with equality, with two sorts, *atom* and *value*. Let *At* stand for the set of atoms in the language of $T$. In $\mathcal{L}$, the set of all object constants of sort *atom* is exactly the set *At*. The symbols *True* and *False* will be the two object constants of sort *value*. $\mathcal{L}$ includes exactly two predicates, in addition to equality: a unary predicate *Holds* of sort *atom* and a binary predicate *Caused* of sort *atom* $\times$ *value*. We will use a variable $x$ of sort *atom*, and a variable $v$ of sort *value*.

We begin the description of $ct(T)$ by letting $\mathcal{C}(T)$ stand for the sentence

$$\bigwedge_{\phi \in T} \mathcal{C}(\phi)$$

where $\mathcal{C}(\phi)$ is defined recursively, as follows.

$$\begin{array}{lll}
\mathcal{C}(p) & = \mathit{Holds}(p) & \text{if } p \in At \\
\mathcal{C}(\mathsf{C}p) & = \mathit{Caused}(p, \mathit{True}) & \text{if } p \in At \\
\mathcal{C}(\mathsf{C}\neg p) & = \mathit{Caused}(p, \mathit{False}) & \text{if } p \in At \\
\mathcal{C}(\top) & = \top & \\
\mathcal{C}(\bot) & = \bot & \\
\mathcal{C}(\neg\phi) & = \neg\mathcal{C}(\phi) & \\
\mathcal{C}((\phi \odot \psi)) & = (\mathcal{C}(\phi) \odot \mathcal{C}(\psi)) &
\end{array}$$

Here $\odot$ stands for any of the binary propositional connectives. Notice that this definition depends on the finiteness of the UCL theory $T$, as well as the assumption that the modal operator $\mathsf{C}$ is applied only to literals. Notice also that $\mathcal{C}(T)$ is ground.

We'll want a unique names axiom (UNA) to say that all object constants of sort *atom* denote distinct domain objects. Thus, UNA stands for the conjunction of all formulas $p \neq q$ such that $p$ and $q$ are distinct members of *At*. Notice that this definition depends on the finiteness of the signature of $T$.

The complete embedding $ct(T)$ consists of the following five sentences.

[5] Familiarity with circumscription will be assumed. See, for example, [29].

$$\mathrm{CIRC}[\mathcal{C}(T) : \mathit{Caused}] \tag{36}$$
$$\forall x\,(\mathit{Holds}(x) \equiv \mathit{Caused}(x, \mathit{True})) \tag{37}$$
$$\forall x\,(\neg \mathit{Holds}(x) \equiv \mathit{Caused}(x, \mathit{False})) \tag{38}$$
$$\mathrm{UNA} \tag{39}$$
$$\forall v (v = \mathit{True} \not\equiv v = \mathit{False}) \tag{40}$$

Notice that second-order quantification is used only implicitly in the embedding $ct(T)$, in (36). The models of (36) are simply the models of $\mathcal{C}(T)$ in which the extent of *Caused* is minimal (for a fixed universe and fixed interpretation of all nonlogical constants except *Caused*).

For every model $M$ of $ct(T)$, there is a one-to-one correspondence between the domain objects of sort *atom* and the members of $At$. To see this, first notice that, because of the UNA, $M$ maps each pair of distinct members of $At$ to distinct domain objects. Now, suppose there is a domain object $\delta$ of sort *atom* such that $M$ maps no member of $At$ to $\delta$. Because $\mathcal{C}(T)$ is ground, and $M$ is a model of $\mathcal{C}(T)$ in which the extent of the predicate *Caused* is minimal, we can conclude that neither $\langle \delta, \mathit{True}^M \rangle$ nor $\langle \delta, \mathit{False}^M \rangle$ belong to $\mathit{Caused}^M$. But the axioms (37) and (38) together imply that

$$\forall x (\mathit{Caused}(x, \mathit{True}) \not\equiv \mathit{Caused}(x, \mathit{False}))\,. \tag{41}$$

Given, in addition, the axiom (40) expressing the unique names and domain closure assumptions for sort *values*, we can conclude that every model of $ct(T)$ is isomorphic to some Herbrand model of $ct(T)$. Thus, in what follows, we restrict our attention to Herbrand interpretations.

For every UCL structure $(I, S)$, let $M(I, S)$ be the Herbrand interpretation of $\mathcal{L}$ such that, for every $p \in At$, the following three conditions hold.

- $M(I, S) \models \mathit{Holds}(p)$ iff $(I, S) \models p$
- $M(I, S) \models \mathit{Caused}(p, \mathit{True})$ iff $(I, S) \models \mathsf{C}p$
- $M(I, S) \models \mathit{Caused}(p, \mathit{False})$ iff $(I, S) \models \mathsf{C}\neg p$

The following lemma is a straightforward consequence of the definitions.

**Lemma 10** *Let $T$ be a finite UCL theory, with finite signature, in which $\mathsf{C}$ is applied only to literals. For any UCL structure $(I, S)$, $(I, S) \models T$ if and only if $M(I, S) \models \mathcal{C}(T)$.*

**Theorem 11** *Let $T$ be a finite UCL theory, with finite signature, in which $\mathsf{C}$ is applied only to literals. An interpretation $I$ is causally explained by $T$ if and only if $M(I, \{I\})$ is a model of $ct(T)$. Moreover, every model of $ct(T)$ is isomorphic to an interpretation $M(I, \{I\})$, for some interpretation $I$ of the language of $T$.*

**Proof.** We've already established that every model of $ct(T)$ is isomorphic to a Herbrand model. In light of (37) and (38), every Herbrand model can be written in the form $M(I, \{I\})$. Now we turn to the first part of the theorem.

($\Longrightarrow$) Assume $I$ is causally explained by $T$. Thus, $(I, \{I\}) \models T$. By Lemma 10, $M(I, \{I\}) \models \mathcal{C}(T)$. Also, $M(I, \{I\})$ clearly satisfies (37)–(40). It remains only to show that the extent of *Caused* in $M(I, \{I\})$ is minimal among models of $\mathcal{C}(T)$ with the same universe, and the same interpretation of all nonlogical constants except *Caused*. Any possible counterexample can be written in the form $M(I, S)$, for some superset $S$ of $\{I\}$. So assume that $M(I, S) \models \mathcal{C}(T)$. By Lemma 10, $(I, S) \models T$. Since $(I, \{I\})$ is the unique $I$-model of $T$, $S = \{I\}$.

($\Longleftarrow$) Assume $M(I, \{I\})$ is a model of $ct(T)$. By Lemma 10, $(I, \{I\}) \models T$. Let $S$ be a superset of $\{I\}$ such that $(I, S) \models T$. By Lemma 10, $M(I, S) \models \mathcal{C}(T)$. Because the extent of *Caused* in $M(I, \{I\})$ is minimal among Herbrand models of $\mathcal{C}(T)$ with the same interpretation of all nonlogical constants except *Caused*, we can conclude that $M(I, S) = M(I, \{I\})$. It follows that $S = \{I\}$. Hence $(I, \{I\})$ is the unique $I$-model of $T$. That is, $I$ is causally explained by $T$.

Notice that any UCL theory of the form allowed by Theorem 11 is S5-equivalent to a UCL theory whose sentences have the form (26) with $\alpha$ a conjunction of literals and all $\beta_i$'s and $\gamma_i$'s literals as well. By Theorem 4, such a UCL theory corresponds to a disjunctive default theory, which itself, because of its "restriction to literals," corresponds to a disjunctive logic program under the answer set semantics, as established in [17]. Hence, given our Theorem 4, the essence of Theorem 11 above can be deduced from Theorem 5.2 (along with subsequent remarks) in [36], where an embedding of disjunctive logic programming into circumscription is described.

## 9 UCL and Lin's Causal Action Theories

Lin [33] recently introduced a causal approach to reasoning about action based on circumscription. In this section, we explore the relationship between Lin's circumscriptive action theories and the UCL action theories described in Section 5, restricted to the case when $\mathsf{C}$ is applied only to literals. We show that on a wide range of action domains, the two approaches coincide.

### *9.1 Lin's Circumscriptive Action Theories*

For the purpose of comparison, we present an account of Lin's proposal that is simplified in several ways. We do not consider non-propositional fluent and action symbols. We also do not employ the situation calculus. Instead we model worlds in which time has the structure of the natural numbers. As discussed in Section 7, this simplifies matters somewhat, eliminating the need for a *Poss* predicate. Finally, we include a domain closure assumption for fluents. In the case of propositional fluents, this is not very significant.

In some other ways, the circumscriptive approach that we describe is more general than Lin's. Because our language includes propositions about the occurrence and non-occurrence of actions, we can accomodate concurrent actions more easily than Lin. We also accomodate a wider variety of causal laws. For instance, we allow formulas expressing causal laws that refer to more than one time point and yet do not involve the occurrence of an action. We allow also for causal laws that involve more than two time points, and we do not require that the time points be successive.

The language of the circumscriptive theory is constructed in the same manner as in the previous section, on the basis of the signature *At* of an underlying propositional language. For this purpose, we employ $\mathcal{L}(\mathbf{F}, \mathbf{A}, \mathbf{T})$ languages, as described in Section 5.1, under the additional restriction that each of the sets $\mathbf{F}$, $\mathbf{A}$, and $\mathbf{T}$ is finite.

For present purposes, we will say that a *Lin formula* is the conjunction of a finite set of ground sentences in which *Caused* appears only positively, and at most once in each sentence.

The next few observations help characterize the relationship between Lin formulas as specified here and the kinds of circumscriptive action theories described in [33]. Assume that $\sigma$ is a fluent designating formula, $A$ is an action name, $F$ is a fluent name, and $V$ is either *True* or *False*. We will write $Holds(\sigma_t)$ to stand for the formula obtained by replacing every occurrence of every fluent atom $f_t$ in $\sigma_t$ by $Holds(f_t)$. Lin's "direct effect" axioms correspond to schemas of the form

$$Holds(A_t) \wedge Holds(\sigma_t) \supset Caused(F_{t+1}, V)\,. \tag{42}$$

Lin's "causal rule" axioms correspond to schemas of the form

$$Holds(\sigma_t) \supset Caused(F_t, V)\,. \tag{43}$$

Lin's "explicit precondition" axioms correspond to schemas of the form

$$Holds(A_t) \supset Holds(\sigma_t)\,.$$

Lin's "qualification state constraint" axioms correspond to schemas of the form

$$Holds(\sigma_t)\,.$$

For example, consider again the Suitcase domain from [33]. We'll use almost the same $\mathcal{L}(\mathbf{F}, \mathbf{A}, \mathbf{T})$ language as in Section 5.2, but restrict time to a finite initial segment of the natural numbers. The Lin formula for this example is characterized by the following schemas of type (42) and (43), where $l$ is again metavariable ranging over $\{L_1, L_2\}$.

$$Holds(Toggle(l)_t) \wedge Holds(Up(l)_t) \supset Caused(Up(l)_{t+1}, False) \quad (44)$$
$$Holds(Toggle(l)_t) \wedge \neg Holds(Up(l)_t) \supset Caused(Up(l)_{t+1}, True) \quad (45)$$
$$Holds(Close_t) \supset Caused(IsOpen_{t+1}, False) \quad (46)$$
$$Holds(Up(L_1)_t) \wedge Holds(Up(L_2)_t) \supset Caused(IsOpen_t, True) \quad (47)$$

Let $T_5$ be the UCL theory given by schemas (9)–(12), which express the domain specific part of the UCL description of the Suitcase domain from Section 5.2. The conjunction of the sentences given by schemas (44)–(47) is exactly $\mathcal{C}(T_5)$, where $\mathcal{C}$ is the translation function defined in Section 8.

Given a Lin formula $D$, the complete circumscriptive action theory $cat(D)$ consists of

$$\text{CIRC}[\,D : Caused\,]$$

along with a number of additional axioms, as follows.

We need axioms expressing domain closure and unique names assumptions for both sorts, which can be written as follows.

$$\forall x \left( \bigvee_{p \in At} x = p \right) \quad (48)$$
$$\bigwedge_{p,q \in At,\, p \neq q} p \neq q \quad (49)$$
$$\forall v (v = True \not\equiv v = False) \quad (50)$$

We also need the following axioms, saying that whatever is caused is true.

$$\forall x (Caused(x, True) \supset Holds(x)) \quad (51)$$
$$\forall x (Caused(x, False) \supset \neg Holds(x)) \quad (52)$$

Finally, we need the inertia axioms given by the schema

$$Holds(f_{t+1}) \equiv (Holds(f_t) \wedge \neg Caused(f_{t+1}, False)) \vee Caused(f_{t+1}, True) \quad (53)$$

where $f$ is a metavariable ranging over fluent names.

Notice the similarity between the circumscriptive action theories $cat(D)$ defined here and the circumscriptive theories $ct(T)$ from Section 8. In both cases, *Caused* is minimized in one part of the theory, and the result is conjoined with several additional axioms. Two differences are especially interesting. Roughly speaking, axioms (37) and (38) of *ct*($T$) are replaced here with the weaker (51) and (52). Intuitively, this is because models of Lin's circumscriptive action theories are not required to satisfy the principle of universal causation. On the other hand, the inertia schema (53) that is built into Lin's theory has no immediate counterpart in *ct*($T$), since inertia is expressed directly in UCL action descriptions (that is, by means of formulas of $T$, which are translated just like any other formulas in $T$).

### *9.2 Lin's Circumscriptive Action Theories in UCL*

The first thing to observe is that, for every Lin formula $D$, there is a UCL theory $T$ in language $\mathcal{L}(\mathbf{F}, \mathbf{A}, \mathbf{T})$ such that $\mathcal{C}(T) = D$. We will show that there is an extension *uclat*($D$) of $T$ such that the interpretations causally explained by *uclat*($D$) correspond to the models of *cat*($D$). We obtain *uclat*($D$) by adding to $T$ the formulas given by the standard schemas (13)–(17) from Section 5.2, taking $\mathbf{I}$ to be the set of all fluent names and their negations. (In light of Theorem 1, any such UCL theory can be expressed as a causal theory. Similarly, by Theorem 3, any such UCL theory has a concise translation into classical propositional logic.)

**Theorem 12** *For any Lin formula $D$, an interpretation $I$ is causally explained by uclat($D$) if and only if there is a superset $S$ of $\{I\}$ such that $M(I, S)$ is a model of cat($D$). Moreover, every model of cat($D$) is isomorphic to an interpretation $M(I, S)$, for some UCL structure $(I, S)$.*

The following easy corollary provides a weaker but more immediate description of the sense in which Lin's circumscriptive causal theories of action are captured in UCL.

**Corollary 13** *For any Lin formula $D$,*

- *for any model $M$ of the corresponding circumscriptive action theory cat($D$), there is an interpretation $I$ causally explained by uclat($D$) such that, for all fluents $f$ and times $t$,*

$$M \models \mathit{Holds}(f_t) \text{ iff } I \models f_t$$

- *for any interpretation $I$ causally explained by uclat($D$), there is a model $M$ of cat($D$) such that, for all fluents $f$ and times $t$,*

$$M \models \mathit{Holds}(f_t) \text{ iff } I \models f_t\,.$$

We begin the proof of Theorem 12 with a straightforward lemma.

**Lemma 14** *Let $T$ be a UCL theory with no nested occurrences of $\mathsf{C}$, in which $\mathsf{C}$ occurs only positively. If $(I, S) \models T$, then for all subsets $S'$ of $S$ such that $I \in S'$, $(I, S') \models T$. If, in addition, $\mathsf{C}$ occurs at most once in each formula, then $(I, S \cup S') \models T$ whenever $(I, S) \models T$ and $(I, S') \models T$.*

**Proof of Theorem 12.** Axioms (48)–(50) allow us to consider only Herbrand models of $cat(D)$. Axioms (51) and (52) show that every Herbrand model of $cat(D)$ can be expressed in the form $M(I, S)$, for some UCL structure $(I, S)$. That proves the second part of the theorem. Now we turn to the first.

Let $T$ be the UCL theory such that $\mathcal{C}(T) = D$. (Here we assume the most natural choice of $T$, which satisfies the conditions of Lemma 14.) Let $T' = uclat(D)$.

($\Longrightarrow$) Assume $I$ is causally explained by $T'$. Thus, $(I, \{I\}) \models T$, and by Lemma 10, $M(I, \{I\}) \models D$. It follows that there is a superset $S$ of $\{I\}$ such that $M(I, S)$ is a model of CIRC[$D : Caused$]. Clearly, $M(I, S)$ satisfies the standard axioms (48)–(52). It remains to show that $M(I, S)$ satisfies the inertia axioms given by (53). Suppose otherwise. Thus, there is a fluent atom $f_t$ such that either $M(I, S) \models \neg Holds(f_t) \wedge Holds(f_{t+1}) \wedge \neg Caused(f_{t+1}, True)$ or $M(I, S) \models Holds(f_t) \wedge \neg Holds(f_{t+1}) \wedge \neg Caused(f_{t+1}, False)$. We now argue the first case. (The second is similar.) By Lemma 10, $(I, S) \models \neg f_t \wedge f_{t+1} \wedge \neg \mathsf{C} f_{t+1}$. Let $I' = I \cup \{\neg f_{t+1}\} \setminus \{f_{t+1}\}$. Notice that $I' \neq I$, with $I \models f_{t+1}$ and $I' \models \neg f_{t+1}$. Since $M(I, S) \not\models Caused(f_{t+1}, True)$ and $I \in S$, we can conclude by choice of $I'$ that $M(I, S) = M(I, S \cup \{I'\})$. Thus, $M(I, S \cup \{I'\}) \models D$, and by Lemma 10, $(I, S \cup \{I'\}) \models T$. By Lemma 14, $(I, \{I, I'\}) \models T$. One easily verifies that $(I, \{I, I'\})$ also satisfies (13)–(17). So $(I, \{I, I'\}) \models T'$, which contradicts the assumption that $I$ is causally explained by $T'$.

($\Longleftarrow$) Assume $M(I, S)$ is a model of $cat(D)$. So $M(I, S) \models D$. By Lemma 10, $(I, S) \models T$. By Lemma 14, $(I, \{I\}) \models T$. One easily verifies that, since $T'$ is obtained from $T$ by adding (13)–(17), $(I, \{I\}) \models T'$. We wish to show that $(I, \{I\})$ is the unique $I$-model of $T'$. Suppose otherwise. So there is a strict superset $S'$ of $\{I\}$ such that $(I, S') \models T'$, and there is a literal $L$ such that $(I, S') \models L \wedge \neg \mathsf{C} L$. In light of (13)–(16), $L$ is a fluent literal that refers to a non-zero time. Assume $L$ has the form $f_{t+1}$. (The argument is analagous if $L$ is $\neg f_{t+1}$.) So $(I, S') \models f_{t+1} \wedge \neg \mathsf{C} f_{t+1}$. In light of (17), $(I, S') \models \neg f_t$. So $(I, S') \models \neg f_t \wedge f_{t+1}$, and thus $(I, S) \models \neg f_t \wedge f_{t+1}$ also. Hence, by Lemma 10, $M(I, S) \models \neg Holds(f_t) \wedge Holds(f_{t+1})$. So by (53), $M(I, S) \models Caused(f_{t+1}, True)$. On the other hand, since $(I, S') \not\models \mathsf{C} f_{t+1}$, $M(I, S') \not\models Caused(f_{t+1}, True)$, again by Lemma 10. Consequently, $M(I, S \cup S') \not\models Caused(f_{t+1}, True)$, which shows that $M(I, S \cup S') \neq M(I, S)$. Since $(I, S') \models T'$, $(I, S') \models T$. Since $(I, S) \models T$ also, we know by Lemma 14 that $(I, S \cup S') \models T$. Hence $M(I, S \cup S') \models D$, once more by Lemma 10. Since $M(I, S)$ is a model of CIRC[$D : Caused$], we can conclude that $M(I, S \cup S') = M(I, S)$, thus reaching contradiction.

### 9.3 Discussion

In [33], Lin briefly discusses the possibility of using a more general form of "causal rule" axiom (43), in which *Caused* can occur negatively any number of times in a sentence, in addition to the one positive occurrence. For example, he suggests extending the circumscriptive action theory for the Suitcase domain with an additional fluent *IsClosed*, understood as the antonym of *IsOpen*, and adding (essentially) the schemas

$$Caused(IsClosed_t, True) \equiv Caused(IsOpen_t, False) \tag{54}$$
$$Caused(IsClosed_t, False) \equiv Caused(IsOpen_t, True) \tag{55}$$

to reflect this understanding. Notice that this resembles the notion of a "defined fluent," discussed in Section 5.3.3, according to which one would augment the UCL Suitcase domain description $T_4$ from Section 5.2 with

$$\mathsf{C}\,(IsClosed_t \equiv \neg IsOpen_t)\ . \tag{56}$$

The first thing to observe is that (56) entails $IsClosed_t \equiv \neg IsOpen_t$ (in S5), while (54) and (55) do not entail

$$Holds(IsClosed_t) \equiv \neg Holds(IsOpen_t)\,. \tag{57}$$

This correctly suggests that some models of the circumscriptive action theory fail to satisfy (57), for some time names $t$.[6] It appears that, in the case of this example, one can obtain a more satisfactory "definition" of *IsClosed* by also including (57) in the circumscriptive theory. In general though, it is unclear how to introduce defined fluents in Lin's circumscriptive action theories.

A related complication arises if we try, for instance, to replace the causal rule axiom (47) with

$$Caused(Up(L_1)_t, True) \wedge Caused(Up(L_2)_t, True) \supset Caused(IsOpen_t, True)$$

(which is closely related to schema (35) from the alternative UCL Suitcase domain formalization in Section 7). This replacement greatly alters the meaning of the circumscriptive action theory. For instance, it allows models that fail to satisfy

$$Holds(Up(L_1)_t) \wedge Holds(Up(L_2)_t) \supset Holds(IsOpen_t)\,.$$

It also makes it impossible, intuitively speaking, to open the suitcase unless one toggles both latches at the same time.

[6] In such models, (57) is false for some initial segment of times names, until, intuitively, some action causes the suitcase to open or close, at which point, roughly speaking, (54) and (55) finally come into play.

In a subsequent paper [34], Lin investigates how to extend his circumscriptive action theories to accomodate nondeterministic actions. For our purposes, the first thing to observe is that nondeterministic actions can typically be described using the natural counterpart to the approach from Section 5.3.2. For instance, one can describe the nondeterministic effect of coin tossing by the schemas

$$Holds(Toss_t) \wedge Holds(Heads_{t+1}) \supset Caused(Heads_{t+1}, True) \qquad (58)$$

$$Holds(Toss_t) \wedge \neg Holds(Heads_{t+1}) \supset Caused(Heads_{t+1}, False) \qquad (59)$$

which correspond to the UCL schemas (18) and (19) from Section 5.3.2. Lin does not (directly) consider this approach. Instead, he begins by considering a variety of methods involving sentences with multiple positive occurrences of *Caused*. For instance, he (essentially) considers a coin-toss axiom like

$$Holds(Toss_t) \supset Caused(Heads_{t+1}, True) \vee Caused(Heads_{t+1}, False)\,. \qquad (60)$$

Notice that, in the presence of standard axioms (51) and (52) guaranteeing that whatever is caused obtains, one can equivalently replace (60) with (58) and (59). In UCL, the corresponding formula

$$Toss_t \supset \mathsf{C}Heads_{t+1} \vee \mathsf{C}\neg Heads_{t+1}$$

also works, since it is S5-equivalent to the conjunction of (18) and (19). But in general such approaches do not translate faithfully into UCL. For instance, if we were to add an action *GetOneUp* to the Suitcase domain, using the UCL schema

$$GetOneUp_t \supset \mathsf{C}\,Up(L_1)_{t+1} \vee \mathsf{C}\,Up(L_2)_{t+1}$$

to describe its effects, the action would never cause the second latch to go up, when performed alone, if the first latch was already up.

Lin shows particular interest in two special cases of nondeterministic effects, which he calls "inclusive" and "exclusive." Inclusive nondeterminism corresponds, in the UCL setting, to families of effect axioms of the following form, where $A$ is an action name, and $\sigma^0, \sigma^1, \ldots, \sigma^n$ are fluent designating formulas.

$$A_t \wedge \sigma^0_t \supset \mathsf{C}\sigma^1_{t+1} \vee \cdots \vee \mathsf{C}\sigma^n_{t+1}$$

$$A_t \wedge \sigma^0_t \supset \mathsf{C}\sigma^1_{t+1} \vee \mathsf{C}\neg\sigma^1_{t+1}$$

$$\vdots$$

$$A_t \wedge \sigma^0_t \supset \mathsf{C}\sigma^n_{t+1} \vee \mathsf{C}\neg\sigma^n_{t+1}$$

The first of these axioms, in the presence of the subsequent axioms, can be equivalently replaced (in S5) by

$$A_t \wedge \sigma^0_t \supset \sigma^1_{t+1} \vee \cdots \vee \sigma^n_{t+1}\,.$$

We can also equivalently replace each of the subsequent axioms with the following pair.

$$A_t \wedge \sigma_t^0 \wedge \sigma_{t+1}^k \supset \mathsf{C}\sigma_{t+1}^k$$
$$A_t \wedge \sigma_t^0 \wedge \neg\sigma_{t+1}^k \supset \mathsf{C}\neg\sigma_{t+1}^k$$

Notice that these transformations yield formulas in which $\mathsf{C}$ occurs at most once, and only positively. Analagous equivalence transformations apply to the corresponding axioms in Lin's theory, given the axioms (51) and (52) guaranteeing that any fluent literal that is caused is true. These observations show that Lin's proposal for inclusive nondeterminism can be applied in the UCL setting, on the basis of Theorem 12. In fact, what we see is that Lin's method for inclusive nondeterminism is essentially a variant of the approach to nondeterminism described briefly in Section 5.3.2. The same observations apply to Lin's proposal for exclusive nondeterminism, which, in the UCL setting, is equivalent to augmenting the inclusive nondeterminism axioms with the additional axiom

$$A_t \wedge \sigma_t^0 \supset \bigwedge_{1 \leq i < j \leq n} \neg(\sigma_{t+1}^i \wedge \sigma_{t+1}^j)\,.$$

Ultimately, Lin [34] introduces an alternative general method for formalizing nondeterminism, using auxiliary *Case* symbols to distinguish between possible nondeterministic outcomes. Without going into details, we note that Lin's "cases" method is easily adapted to the UCL setting.

Finally, while Theorem 12 embeds Lin's circumscriptive action theories in UCL, it is interesting to consider what happens when we proceed in the opposite direction. Let us assume that a UCL theory $T$ is finite, with a finite signature, and that $\mathsf{C}$ is applied only to literals. In this case, the translation $\mathcal{C}(T)$ is defined. Assume in addition that $T$ includes the formulas given by the standard schemas (13)–(17), with $\mathbf{I}$ consisting of the fluent names and their negations. If we extend the definition of *cat* so that it applies even when *Caused* is allowed to occur negatively and more than once in each sentence, then it is straightforward to verify that the circumscriptive theory $cat(\mathcal{C}(T))$ is equivalent to $ct(T)$. In light of Theorem 11, this observation shows that if we augment Lin's action theories with axioms corresponding to the standard schemas, all models satisfy the principle of universal causation, and his approach converges with ours.

In establishing these relationships between Lin's circumscriptive action theories and UCL action theories, we assume that $\mathsf{C}$ is applied only to literals. By doing so, we forfeit the ability to express directly the fact that a complex formula is caused to hold, and so can no longer express traditional ramification constraints or introduce defined fluents by the methods described in Sections 5.3.1 and 5.3.3.

The results in this section also require that the set of inertial fluents be given by the set of fluent names and their negations. This assumption is built into Lin's approach, whereas in UCL action theories in general, as remarked earlier, the inertial fluents may be chosen differently. (This facility too is critical to our method for introducing defined fluents.) In fact, as illustrated in Section 5.3.5, it is possible in UCL to generalize the commonsense law of inertia so as to allow for fluents that tend to change in particular ways (instead of tending to stay the same).

## 10 UCL and Autoepistemic Logic

It may be interesting to consider the mathematical relationship of UCL to autoepistemic logic (AEL), which is surely the most widely-familiar modal nonmonotonic logic. For this purpose we employ the elegant model-theoretic characterization of autoepistemic logic from [32].

Let $T$ be an autoepistemic theory. We say that a set $S$ of interpretations is an *AE model* of $T$ if

$$S = \{I : (I, S) \models T\} .$$

Recall that for autoepistemic logic we do not require that structures $(I, S)$ satisfy the condition $I \in S$.

The definition of an AE model can be reformulated as follows. A set $S$ of interpretations is an AE model of an AEL theory $T$ if and only if, for all interpretations $I$,

$$(I, S) \models T \text{ iff } I \in S . \tag{61}$$

In this form, we can observe a strong resemblance to the fixpoint condition in UCL, which can be similarly reformulated, as follows. An interpretation $I$ is causally explained by a UCL theory $T$ if and only if, for every set $S$ of interpretations such that $I \in S$,

$$(I, S) \models T \text{ iff } S = \{I\} . \tag{62}$$

Roughly speaking, the reversal of the roles of $S$ and $I$ in the fixpoint conditions (61) and (62) is reflected in a corresponding reversal of the role of the modal operator in the two logics. In accordance with this observation, it is not difficult to establish the following. [7]

[7] We will use the symbol $\mathsf{B}$ for the AEL modal operator, rather than $\mathbf{L}$, which is also often used.

**Proposition 15** *Let $T$ be a UCL theory consisting of formulas of the form*

$$\phi \vee \mathsf{C}\psi \tag{63}$$

*where $\phi$ and $\psi$ are nonmodal formulas. Take the AEL theory $T'$ obtained by replacing each UCL formula* (63) *with the AEL formula*

$$\mathsf{B}\phi \vee \psi\,.$$

*An interpretation $I$ is causally explained by $T$ if and only if $\{I\}$ is an AE model of $T'$.*

The translation in Proposition 15, while easy to justify directly, can also be obtained indirectly from Konolige's well-known partial translation of default logic into autoepistemic logic [28], via Theorem 4.

We can obtain a more general result of this kind by using a more complex translation, in which "caused" becomes, roughly speaking, "truly believed."

**Proposition 16** *Let $T$ be a UCL theory consisting of formulas of the form*

$$\phi \vee \mathsf{C}\psi^1 \vee \cdots \vee \mathsf{C}\psi^n \tag{64}$$

*where $\phi, \psi^1, \ldots, \psi^n$ are nonmodal formulas. Take the AEL theory $T'$ obtained by replacing each UCL formula* (64) *with the AEL formula*

$$\mathsf{B}\phi \vee (\psi^1 \wedge \mathsf{B}\psi^1) \vee \cdots \vee (\psi^n \wedge \mathsf{B}\psi^n)\,. \tag{65}$$

*An interpretation $I$ is causally explained by $T$ if and only if $\{I\}$ is an AE model of $T'$.*

**Proof.** ($\Longrightarrow$) Assume that $I$ is causally explained by $T$. So $(I, \{I\}) \models T$, and it follows easily that $(I, \{I\}) \models T'$. Let $I'$ be such that $(I', \{I\}) \models T'$. So for any formula (65) in $T'$, either $I \models \phi$ or both $I$ and $I'$ satisfy one of $\psi^1, \ldots, \psi^n$. It follows that either $(I, \{I, I'\}) \models \phi$ or $(I, \{I, I'\})$ satisfies one of $\mathsf{C}\psi^1, \ldots, \mathsf{C}\psi^n$. Hence $(I, \{I, I'\}) \models T$, and since $(I, \{I\})$ is the unique $I$-model of $T$, $I' = I$. Consequently, $\{I\}$ is an AE model of $T'$.

($\Longleftarrow$) Assume that $\{I\}$ is an AE model of $T'$. So $(I, \{I\}) \models T'$, and it follows easily that $(I, \{I\}) \models T$. Let $I'$ be such that $(I, \{I, I'\}) \models T$. So for any formula (64) in $T$, either $I \models \phi$ or both $I$ and $I'$ satisfy one of $\psi^1, \ldots, \psi^n$. Therefore either $(I', \{I\}) \models \mathsf{B}\phi$ or $(I', \{I\})$ satisfies one of $\psi^1 \wedge \mathsf{B}\psi^1, \ldots, \psi^n \wedge \mathsf{B}\psi^n$. Hence $(I', \{I\}) \models T'$. Since $\{I\}$ is an AE model of $T'$, we can conclude that $I' = I$. Consequently, $I$ is causally explained by $T$.

The natural extension of Proposition 16 to formulas with negative occurrences of $\mathsf{C}$ is not sound. For example, the UCL theory

$$\mathsf{C}(p \vee q)$$
$$\mathsf{C}p \supset \mathsf{C}q$$
$$\mathsf{C}q \supset \mathsf{C}p$$

has no causally explained interpretations, while $\{\{p, q\}\}$ is an AE model of

$$(p \vee q) \wedge \mathsf{B}(p \vee q)$$
$$p \wedge \mathsf{B}p \supset q \wedge \mathsf{B}q$$
$$q \wedge \mathsf{B}q \supset p \wedge \mathsf{B}p\,.$$

This approach works though if $\mathsf{C}$ is applied only to literals. In fact, the correctness of the resulting translation can be deduced from an embedding of disjunctive logic programming under the answer set semantics into autoepistemic logic proposed simultaneously in [6,32,37], again via Theorem 4.

For our purposes, we can formulate the translation as follows. If $\mathsf{C}$ is applied only to literals in a UCL formula, then the following translation is defined.

$$\begin{array}{lll} \mathcal{A}(\phi) & = \mathsf{B}\phi & \text{if } \phi \text{ is nonmodal} \\ \mathcal{A}(\mathsf{C}L) & = (L \wedge \mathsf{B}L) & \text{if } L \text{ is a literal} \\ \mathcal{A}(\neg\phi) & = \neg\mathcal{A}(\phi) & \text{if } \phi \text{ is not nonmodal} \\ \mathcal{A}((\phi \odot \psi)) & = (\mathcal{A}(\phi) \odot \mathcal{A}(\psi)) & \text{if } \phi \text{ or } \psi \text{ is not nonmodal} \end{array}$$

Here $\odot$ stands for any of the binary propositional connectives.

**Proposition 17** *Let $T$ be a UCL theory in which $\mathsf{C}$ is applied only to literals. Take the AEL theory $T'$ obtained by replacing each UCL formula $\phi$ with the AEL formula $\mathcal{A}(\phi)$. An interpretation $I$ is causally explained by $T$ if and only if $\{I\}$ is an AE model of $T'$.*

It is unclear what lessons to draw from such mathematical facts. Notice that, for AE models of the form $\{I\}$, the fixpoint condition involves only structures of the form $(I', \{I\})$. Therefore, one can, for instance, replace $\mathsf{B}\phi$ with $\neg\mathsf{B}\neg\phi$ without affecting the "complete" AE models. Similarly, one can replace $\mathsf{B}(\phi \vee \psi)$ with $\mathsf{B}\phi \vee \mathsf{B}\psi$. These observations demonstrate that the "complete" subset of autoepistemic logic is relatively inexpressive as a logic of belief, as one would intuitively expect. Nonetheless, Propositions 15–17 show that interesting causal theories of action can be captured in this subset of AEL.

## 11 Nonpropositional UCL

The signature of a (nonpropositional) UCL language is given by a set of non-logical constants: that is, function symbols (with arities and sorts) and predicate symbols (with arities and sorts). (Object constants are functions of arity zero, and propositional constants are predicates of arity zero.) The definitions of a formula, sentence, theory, free occurrence of a variable and so on are as expected.

A UCL *structure* is a pair $(I, S)$ such that $S$ is a set of interpretations with the same universe, and $I \in S$.

In the recursive truth definition, we extend the language each time a quantifier is encountered, adding a new nonlogical constant of the appropriate sort. To this end, we introduce the following auxiliary definition.

Let $(I, S)$ be a UCL structure for a given language. When we add a new nonlogical constant $X$ to the signature, we call a UCL structure $(I', S')$ for the resulting language an *$X$-extension* of $(I, S)$ if $I'$ is an extension of $I$ and $S'$ is obtained by taking, for each member of $S$, the unique extension that interprets $X$ as $I'$ does.

The truth of a UCL sentence in a UCL structure is defined by the standard recursions over the propositional connectives, plus the following four cases. (We assume in the last two cases that $X$ is a new nonlogical constant of the same sort as the variable $x$, and by $\phi(X)$ we denote the formula obtained from $\phi(x)$ by simultaneously replacing each free occurrence of $x$ by $X$.)

$$
\begin{array}{lll}
(I, S) \models P & \text{iff } I \models P & \text{(for any ground atom } P\text{)} \\
(I, S) \models \mathsf{C}\phi & \text{iff for all } I' \in S,\ (I', S) \models \phi & \\
(I, S) \models \forall x \phi(x) & \text{iff for every } X\text{-extension } (I', S'),\ (I', S') \models \phi(X) & \\
(I, S) \models \exists x \phi(x) & \text{iff for some } X\text{-extension } (I', S'),\ (I', S') \models \phi(X) &
\end{array}
$$

It is often convenient to designate some nonlogical constants *exempt*, which, intuitively, exempts them from the principle of universal causation. Mathematically, this practice is reflected in the definition of an *$I$-structure*: a UCL structure $(I, S)$ such that all members of $S$ interpret all exempt symbols exactly as $I$ does. An *$I$-model* of a UCL theory $T$ is an $I$-structure that is a model of $T$.

The definition of a causally explained interpretation is just as it was in the propositional case. An interpretation $I$ is *causally explained* by a UCL theory $T$ if $(I, \{I\})$ is the unique $I$-model of $T$.

Clearly, this definition extends the definition introduced in Section 2 for propositional UCL, assuming that the nonlogical constants of the language (i.e. the propositional symbols) are not declared exempt.

In fact, the introduction of exempt nonlogical constants, although convenient, is not mathematically necessary. That is, as in the propositional case, we can effectively exempt a nonlogical constant from the principle of universal causation by writing appropriate axioms. If $P$ is a predicate symbol, and $\overline{x}$ a suitable tuple of variables, we can make $P$ exogenous, essentially as before, by adding the pair of axioms

$$\forall \overline{x}\, (P(\overline{x}) \supset \mathsf{C}P(\overline{x})) \tag{66}$$

$$\forall \overline{x}\, (\neg P(\overline{x}) \supset \mathsf{C}\neg P(\overline{x}))\ . \tag{67}$$

To see this, notice that, for any suitable choice of signature, for any interpretation $I$, the $I$-models of (66)–(67) are exactly the $I$-structures obtained if $P$ is designated exempt. That is, any $I$-structure $(I, S)$ that satisfies (66) and (67) will have the property that every element of $S$ interprets $P$ exactly as $I$ does, and conversely, (66) and (67) are satisfied by any $I$-structure $(I, S)$ in which every element of $S$ interprets $P$ exactly as $I$ does. A similar approach works for function symbols. If $F$ is a function symbol, $\overline{x}$ a suitable tuple of variables and $v$ a suitable variable, then we can make $F$ exogenous by writing

$$\forall \overline{x}, v\, (F(\overline{x}) = v \supset \mathsf{C}F(\overline{x}) = v)\ .$$

As an example of the use of nonpropositional UCL, another version of Lin's Suitcase domain is displayed in Figure 1. The signature of the language and the sorts of variables should be clear from context. All nonlogical constants except *Up* and *IsOpen* are exempt. In this version of the Suitcase domain, the time structure is axiomatized. (We abbreviate $\mathit{Successor}(t)$ as $t'$.)

## 12 Nonpropositional Causal Theories in UCL

Here we review Lifschitz's definition of nonpropositional causal theories [30], altering some terminology and notation to maintain greater consistency with the presentation of (propositional) causal theories in Section 4. We then sketch a proof of the fact that nonpropositional causal theories are subsumed by (nonpropositional) UCL.

As a consequence, UCL inherits relevant results from nonpropositional causal theories. For instance, in [31] Lifschitz gives a situation calculus axiomatization of the blocks world as a first-order causal theory, and relates it to a "Lin-style" formalization, thus providing another perspective on the use of the situation calculus in UCL. In [19], Giunchiglia and Lifschitz define a high-level action

$$\forall t(0 \neq t') \qquad \forall t, u(t' = u' \supset t = u)$$

$$\forall p\,(p(0) \wedge \forall t(p(t) \supset p(t')) \supset \forall t(p(t))) \qquad \forall l(l = L_1 \not\equiv l = L_2)$$

$$\forall l(\mathit{Up}(l,0) \supset \mathsf{C}\,\mathit{Up}(l,0)) \qquad \forall l(\neg\,\mathit{Up}(l,0) \supset \mathsf{C}\neg\,\mathit{Up}(l,0))$$

$$\mathit{IsOpen}(0) \supset \mathsf{C}\mathit{IsOpen}(0) \qquad \neg\mathit{IsOpen}(0) \supset \mathsf{C}\neg\mathit{IsOpen}(0)$$

$$\forall l,t(\mathit{Up}(l,t) \wedge \mathit{Up}(l,t') \supset \mathsf{C}\,\mathit{Up}(l,t'))$$

$$\forall l,t(\neg\,\mathit{Up}(l,t) \wedge \neg\,\mathit{Up}(l,t') \supset \mathsf{C}\neg\,\mathit{Up}(l,t'))$$

$$\forall t(\mathit{IsOpen}(t) \wedge \mathit{IsOpen}(t') \supset \mathsf{C}\mathit{IsOpen}(t'))$$

$$\forall t(\neg\mathit{IsOpen}(t) \wedge \neg\mathit{IsOpen}(t') \supset \mathsf{C}\neg\mathit{IsOpen}(t'))$$

$$\forall l,t\,(\mathit{Toggle}(l,t) \wedge \mathit{Up}(l,t) \supset \mathsf{C}\neg\,\mathit{Up}(l,t'))$$

$$\forall l,t\,(\mathit{Toggle}(l,t) \wedge \neg\,\mathit{Up}(l,t) \supset \mathsf{C}\,\mathit{Up}(l,t'))$$

$$\forall t\,(\mathit{Close}(t) \supset \mathsf{C}\neg\mathit{IsOpen}(t')) \qquad \forall t\,(\forall l(\mathit{Up}(l,t)) \supset \mathsf{C}\mathit{IsOpen}(t))$$

Fig. 1. Lin's Suitcase domain in second-order UCL

language $\mathcal{C}$ based on nonpropositional causal theories, and relate $\mathcal{C}$ to the action language $\mathcal{A}_C$ of Baral and Gelfond [3]. [19] also includes an interesting general theorem relating $\mathcal{C}$ to Lin's circumscriptive causal action theories, providing yet another perspective on the situation calculus in causal theories. (The theorem does not, however, address the question of adequately defining the *Poss* predicate, as discussed briefly in Section 7.) Finally, in [30] Lifschitz extends the literal completion method of [39] to ("definite") nonpropositional causal theories, thus showing how, for example, some first-order UCL theories can be translated into first-order classical logic.

### *12.1 Lifschitz's Nonpropositional Causal Theories*

Begin with a language of classical logic. As in the previous section, some nonlogical constants may be designated exempt. In fact, here we must require that only a finite number of nonlogical constants are not designated exempt. A *causal law* is an expression of the form

$$\phi \Rightarrow \psi$$

where $\phi$ and $\psi$ are formulas of the language. A *causal theory* is a finite set of causal laws. (Notice that, except for the finiteness requirements, this definition of a causal theory extends that of Section 4.) In Lifschitz's account, an interpretation is causally explained by a causal theory just in case it is a model of an associated theory of classical logic, described next.

In what follows, let $\overline{N}$ be a list of all nonexempt nonlogical constants. We say that a list of nonlogical constants or variables is *similar* to $\overline{N}$ if it has the same length as $\overline{N}$ and each of its members is of the same sort as the corresponding member of $\overline{N}$. We can denote a formula (in which none, some, or all nonexempt nonlogical constants appear) by $\phi(\overline{N})$. Then for any list $\overline{M}$ that is similar to $\overline{N}$, we can write $\phi(\overline{M})$ to denote the formula obtained by simultaneously replacing each occurrence of each member of $\overline{N}$ by the corresponding member of $\overline{M}$.

Consider a nonpropositional causal theory $D$ with causal laws

$$\begin{array}{c}\phi_1(\overline{N},x^1) \Rightarrow \psi_1(\overline{N},x^1)\\ \vdots \\ \phi_k(\overline{N},x^k) \Rightarrow \psi_k(\overline{N},x^k)\end{array}$$

where $x^i$ is the list of all free variables for the $i$-th causal law. Let $\overline{n}$ be a list of new variables that is similar to $\overline{N}$. By $D^*(\overline{n})$ we denote the formula

$$\bigwedge_{1 \leq i \leq k} \forall x^i \left( \phi_i(\overline{N},x^i) \supset \psi_i(\overline{n},x^i) \right) .$$

An interpretation is *causally explained* by $D$ if it is a model of

$$\forall \overline{n} \left( D^*(\overline{n}) \equiv \overline{n} = \overline{N} \right) \tag{68}$$

where $\overline{n} = \overline{N}$ stands for the conjunction of the equalities between members of $\overline{n}$ and the corresponding members of $\overline{N}$.

As shown in [30], this definition of a causally explained interpretation extends the definition for propositional causal theories reviewed in Section 4. Notice that for propositional causal theories the corresponding sentence (68) is a quantified boolean formula, from which quantifiers can be eliminated (with worst case exponential increase in size). Thus Lifschitz's definition yields a general translation of propositional causal theories into classical propositional logic.

### *12.2 Nonpropositional Causal Theories in UCL*

Here we sketch a proof of the following theorem.

**Theorem 18** *Let $D$ be a nonpropositional causal theory and let $T$ be the UCL theory obtained by replacing each causal law $\phi \Rightarrow \psi$ by the universal closure of the UCL formula $\phi \supset \mathsf{C}\psi$. An interpretation is causally explained by $D$ if and only if it is causally explained by $T$.*

**Proof sketch.** Let us write $\forall \overline{n} T^*(\overline{n})$ to denote the sentence (68) whose models are the interpretations causally explained by $D$. Extend the language of $\forall \overline{n} T^*(\overline{n})$ as follows. For every $X \in \overline{N}$, add a new nonlogical constant $X'$ of the same sort. Let $\overline{N'}$ be the list of these new symbols, which is similar to $\overline{N}$. Given an interpretation $I$ of the original language, an interpretation $J$ of the new language is called an $I$-interpretation if $J$ extends $I$. (That is, if $J$ has the same universe as $I$ and interprets all nonlogical constants in the original language exactly as $I$ does.) The first observation is that $I \models \forall \overline{n} T^*(\overline{n})$ iff, for every $I$-interpretation $J$, $J \models T^*(\overline{N'})$. Let $\hat{I}$ be the unique $I$-interpretation such that for every $X \in \overline{N}$, $(X')^J = X^I$. The sentence $T^*(\overline{N'})$ is an equivalence whose right-hand side is the sentence $\overline{N'} = \overline{N}$. Since $\hat{I}$ is the only $I$-interpretation that satisfies $\overline{N'} = \overline{N}$, it follows from the previous observation that $I \models \forall \overline{n} T^*(\overline{n})$ iff $\hat{I}$ is the unique $I$-interpretation satisfying $D^*(\overline{N'})$.

The proof can be completed by showing that $\hat{I}$ is the unique $I$-interpretation satisfying $D^*(\overline{N'})$ iff $(I, \{I\})$ is the unique $I$-model of $T$. The first step is to show that $\hat{I} \models D^*(\overline{N'})$ iff $(I, \{I\}) \models T$, by showing that, for any $i$,

$$\begin{aligned} \hat{I} \models \forall x^i \left(\phi_i(\overline{N}, x^i) \supset \psi_i(\overline{N'}, x^i)\right) \text{ iff } & I \models \forall x^i \left(\phi_i(\overline{N}, x^i) \supset \psi_i(\overline{N}, x^i)\right) \\ \text{iff } & (I, \{I\}) \models \forall x^i \left(\phi_i(\overline{N}, x^i) \supset \mathsf{C}\psi_i(\overline{N}, x^i)\right). \end{aligned}$$

All that remains is to prove that if $\hat{I} \models D^*(\overline{N'})$, then some $I$-interpretation other than $\hat{I}$ satisfies $D^*(\overline{N'})$ iff there is a proper superset $S$ of $I$ such that $(I, S) \models T$.

Now we describe observations and a lemma sufficient to complete this last step. First, because $\mathsf{C}$ occurs only positively in $T$, we know that if $(I, S) \models T$, then for any subset $S'$ of $S$ such that $I \in S$, $(I, S') \models T$. Consequently, one need only consider $I$-structures of the form $(I, \{I, I'\})$ in order to determine whether $(I, \{I\})$ is the unique $I$-model of $T$. This is convenient because there is a natural one-to-one correspondence between $I$-interpretations and $I$-structures of the form $(I, \{I, I'\})$ such that, for every $I$-interpretation $J$ and corresponding $I$-structure $(I, \{I, I'\})$, for all $X \in \overline{N}$, $(X')^J = X^{I'}$.

In light of the preceding observations, it remains only to prove the following lemma. If $(I, \{I\}) \models T$, then for any $I$-interpretation $J$ and corresponding $I$-structure $(I, \{I, I'\})$, $J \models D^*(\overline{N'})$ iff $(I, \{I, I'\}) \models T$. For the proof of this lemma, it is convenient to extend the truth definition for UCL to structures of the form $(I, \{I'\})$, where $I'$ may differ from $I$. Under this extended truth definition, one can show that, for any $i$, $J \models \forall x^i \left(\phi_i(\overline{N}, x^i) \supset \psi_i(\overline{N'}, x^i)\right)$ iff $(I, \{I'\}) \models \forall x^i \left(\phi_i(\overline{N}, x^i) \supset \mathsf{C}\psi_i(\overline{N}, x^i)\right)$. So $J \models D^*(\overline{N'})$ iff $(I, \{I'\}) \models T$. To complete the proof of the lemma, notice that $(I, \{I, I'\}) \models T$ iff both $(I, \{I\}) \models T$ and $(I, \{I'\}) \models T$, since $\mathsf{C}$ appears at most once, and only positively, in each sentence of $T$.

## 13 Concluding Remarks

This note has introduced UCL, a modal nonmonotonic logic designed for representing commonsense knowledge about actions. More specifically, in UCL one represents the conditions under which facts are caused. On the basis of this mathematically simple form of causal knowledge, UCL characterizes the worlds that are causally possible. The logic takes its name from the principle of universal causation, the simplifying assumption that underlies the fixpoint definition of a causally explained interpretation.

In applications of UCL to reasoning about action—discussed primarily in Sections 5, 7, and 9—we have seen that universal causation is easily relaxed by means of standard axioms. Alternatively, one can declare a subset of the nonlogical constants exempt from universal causation, as is done in the formalization of the Suitcase domain displayed in Section 11.

The principle of universal causation plays a key role in the simple, robust solution to the frame problem employed throughout the paper. We have seen (in Section 5.3.5) that essentially the same approach can be used to describe inertia in worlds in which, intuitively speaking, things change (in a certain way) unless they are made not to. It is also interesting to note that it appears straightforward to generalize this approach in another direction, to allow for a dense time structure, in which the inertia of a fluent $F$ might be expressed by writing

$$\forall t\, (\exists r\, (r < t \wedge \forall s\, (r \leq s \leq t \supset F(s))) \supset \mathsf{C}F(t)) \ .$$

We have related UCL to Reiter's default logic (and, more generally, disjunctive default logic), circumscription, and autoepistemic logic—in Sections 6, 8, and 10. We showed in Section 4 that UCL extends the causal theories formalism of McCain and Turner, and, in doing so, provides a more adequate semantic account of it. In Section 12, we showed that (nonpropositional) UCL extends the nonpropositional causal theories of Lifschitz.

We have shown that UCL inherits a variety of causal theories of action and change previously proposed in the literature, including in large part the circumscriptive causal action theories of Lin. We have also established, by means of Theorems 1 and 12, the remarkable similarity between the causal action theories of Lin and those of McCain and Turner. In light of this, Theorem 4 and Propositions 15–17 show how such causal action theories can also be expressed in default logic and in autoepistemic logic.

It would be interesting to investigate embeddings in UCL of some of the many other causal approaches to reasoning about action that have been proposed in recent years. Such results can clarify the relationships between the various proposals. They can also help guide future work exploring the range of action domains expressible in UCL.

As a final remark, although we have not explored the general problem of computing causally explained interpretations and UCL entailment, we have noted that a useful subset of UCL corresponds to logic programming under the answer set semantics, which is computed by several fast improving, publicly available systems. We have also identified the class of definite UCL theories, which can be translated into classical propositional theories of comparable size, by the method of literal completion (Section 4.3). For definite UCL theories, automated query answering can be carried out simply by using standard satisfiability tools on the classical propositional theory obtained by literal completion. This idea is implemented in the system CCALC of Norman McCain, which is currently available at `www.cs.utexas.edu/users/mccain/cc`. CCALC can also accept as input action descriptions in the high-level action language $\mathcal{C}$, mentioned previously. A theoretical framework for automated planning on the basis of definite UCL theories is provided by [40], based in part on Kautz and Selman's notion of satisfiability planning [26]. The experimental results reported in [40] demonstrate that this approach can yield relatively good computational performance on classical planning problems.


## Acknowledgements

I am happily indebted to Norm McCain for years of conversation and collaboration on topics related to this paper. I thank participants of the Mini-Workshop on Causality (UT Austin, May 5–6, 1996) for challenging discussions. I am deeply grateful to my PhD advisor, Vladimir Lifschitz, for generous guidance and encouragement. Thanks to Enrico Giunchiglia for timely advice and criticism, and to the anonymous referees for helpful suggestions. This work was partially supported by National Science Foundation Grant IRI-9306751 and University of Minnesota Grant-in-aid #17831.


## References


[1] Andrew Baker. Nonmonotonic reasoning in the framework of situation calculus. *Artificial Intelligence*, 49:5–23, 1991.

[2] Chitta Baral. Reasoning about actions: non-deterministic effects, constraints, and qualifications. In *Proc. of IJCAI-95*, pages 2017–2023, 1995.

[3] Chitta Baral and Michael Gelfond. Representing concurrent actions in extended logic programming. In *Proc. of IJCAI-93*, pages 866–871, 1993.

[4] Gerhard Brewka and Joachim Hertzberg. How to do things with worlds: On formalizing actions and plans. *Journal of Logic and Computation*, 3(5):517–532, 1993.

[5] Marcos Castillo, Olivier Gasquet, and Andreas Herzig. Formalizing action and change in modal logic I: the frame problem. *Journal of Logic and Computation*, 1999. To appear.

[6] Jinhua Chen. Minimal knowledge + negation as failure = only knowing (sometimes). In Luis Moniz Pereira and Anil Nerode, editors, *Logic Programming and Non-monotonic Reasoning: Proceedings of the Second International Workshop*, pages 132–150, 1993.

[7] Pawel Cholewiński, Wiktor Marek, Artur Mikitiuk, and Miroslaw Truszczyński. Computing with default logic. In *Artificial Intelligence*, 1999. To appear.

[8] Keith Clark. Negation as failure. In Herve Gallaire and Jack Minker, editors, *Logic and Data Bases*, pages 293–322. Plenum Press, New York, 1978.

[9] Thomas Eiter, Nicola Leone, Cristinel Mateis, Gerald Pfeifer, and Francesco Scarcello. The KR system DLV: Progress report, comparisons and benchmarks. In *Proc. Sixth Int'l Conf. on Principles of Knowledge Representation and Reasoning*, pages 406–417, 1998.

[10] Charles Elkan. Reasoning about action in first-order logic. In *Proc. of the 1992 Canadian Conf. on Artificial Intelligence*, 1992.

[11] Hector Geffner. *Default Reasoning: Causal and Conditional Theories.* PhD thesis, UCLA, Department of Computer Science, 1989.

[12] Hector Geffner. Causal theories of nonmonotonic reasoning. In *Proc. of AAAI-90*, pages 524–530, 1990.

[13] Hector Geffner. *Reasoning with defaults: causal and conditional theories.* MIT Press, Cambridge, MA, 1992.

[14] Michael Gelfond and Vladimir Lifschitz. Classical negation in logic programs and disjunctive databases. *New Generation Computing*, 9:365–385, 1991.

[15] Michael Gelfond and Vladimir Lifschitz. Representing action and change by logic programs. *Journal of Logic Programming*, 17:301–322, 1993.

[16] Michael Gelfond and Vladimir Lifschitz. Action languages. In *Electronic Transactions on Artificial Intelligence*, 1999.

[17] Michael Gelfond, Vladimir Lifschitz, Halina Przymusińska, and Miroslaw Truszczyński. Disjunctive defaults. In James Allen, Richard Fikes, and Erik Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proc. of the 2nd Int'l Conference*, pages 230–237, 1991.

[18] Laura Giordano, Alberto Martelli, and Camilla Schwind. Dealing with concurrent actions in modal action logic. In *Proc. of the 13th European Conf. on Artificial Intelligence*, pages 537–541, 1998.

[19] Enrico Giunchiglia and Vladimir Lifschitz. An action language based on causal explanation: Preliminary report. In *Proc. AAAI-98*, pages 623–630, 1998.

[20] Enrico Giunchiglia and Vladimir Lifschitz. An action language based on causal logic. In Working Notes of 4th Symp. on Logical Formalizations of Commonsense Reasoning, 1998.

[21] Peter Grünwald. Causation and nonmonotonic temporal reasoning. In G. Brewka, Habel C., and W. Nebel, editors, *KI-97: Advances in Artificial Intelligence (Lecture Notes in Artificial Intelligence no. 1303)*, pages 159–170. Springer, 1997.

[22] J. Gustaffson and P. Doherty. Embracing occlusion in specifying the indirect effects of actions. In *Principles of Knowledge Representation and Reasoning: Proc. of the Fifth Int'l Conference*, 1996.

[23] Steve Hanks and Drew McDermott. Nonmonotonic logic and temporal projection. *Artificial Intelligence*, 33(3):379–412, 1987.

[24] G.E. Hughes and M.J. Cresswell. *An introduction to modal logic.* Methuen and Co LTD, 1968.

[25] G. Neelakantan Kartha and Vladimir Lifschitz. Actions with indirect effects (preliminary report). In *Proc. of the Fourth Int'l Conf. on Principles of Knowledge Representation and Reasoning*, pages 341–350, 1994.

[26] Henry Kautz and Bart Selman. Planning as satisfiability. In J. Lloyd, editor, *Proceedings of the 10th European Conference on Artificial Intelligence (ECAI 92)*, pages 359–379, Vienna, Austria, 1992.

[27] Henry Kautz and Bart Selman. Pushing the envelope: planning, propositional logic, and stochastic search. In *Proceedings of AAAI-96*, pages 1194–1201, 1996.

[28] Kurt Konolige. On the relation between default and autoepistemic logic. *Artificial Intelligence*, 35:343–382, 1988.

[29] Vladimir Lifschitz. Circumscription. In D.M. Gabbay, C.J. Hogger, and J.A. Robinson, editors, *The Handbook of Logic in AI and Logic Programming*, volume 3, pages 298–352. Oxford University Press, 1993.

[30] Vladimir Lifschitz. On the logic of causal explanation. *Artificial Intelligence*, 96:451–465, 1997.

[31] Vladimir Lifschitz. Situation calculus and causal logic. In *Proc. of the Sixth Int'l Conf. on Principles of Knowledge Representation and Reasoning*, pages 536–546, 1998.

[32] Vladimir Lifschitz and Grigori Schwarz. Extended logic programs as autoepistemic theories. In Luis Moniz Pereira and Anil Nerode, editors, *Logic Programming and Non-monotonic Reasoning: Proceedings of the Second Int'l Workshop*, pages 101–114, 1993.

[33] Fangzhen Lin. Embracing causality in specifying the indirect effects of actions. In *Proc. of IJCAI-95*, pages 1985–1991, 1995.

[34] Fangzhen Lin. Nondeterminism in causal theories of action. In *Proc. of AAAI-96*, pages 670–676, 1996.

[35] Fangzhen Lin and Raymond Reiter. State constraints revisited. *Journal of Logic and Computation*, 4(5):655–678, 1994. Special Issue on Actions and Processes.

[36] Fangzhen Lin and Yoav Shoham. A logic of knowledge and justified assumption. *Artificial Intelligence*, 57(2–3):271–290, 1992.

[37] Wiktor Marek and Miroslaw Truszczyński. Reflexive autoepistemic logic and logic programming. In Luis Moniz Pereira and Anil Nerode, editors, *Logic Programming and Non-monotonic Reasoning: Proceedings of the Second International Workshop*, pages 115–131, 1993.

[38] Norman McCain and Hudson Turner. A causal theory of ramifications and qualifications. In *Proc. of IJCAI-95*, pages 1978–1984, 1995.

[39] Norman McCain and Hudson Turner. Causal theories of action and change. In *Proc. of AAAI-97*, pages 460–465, 1997.

[40] Norman McCain and Hudson Turner. Satisfiability planning with causal theories. In *Principles of Knowledge Representation and Reasoning: Proc. of the Sixth Int'l Conference*, 1998.

[41] John McCarthy. *Formalizing common sense: papers by John McCarthy.* Ablex, Norwood, NJ, 1990.

[42] John McCarthy and Patrick Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence*, volume 4, pages 463–502. Edinburgh University Press, Edinburgh, 1969. Reproduced in [41].

[43] Robert Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25(1):75–94, 1985.

[44] Ilkka Niemelä and Patrik Simons. Efficient implementation of the well-founded and stable model semantics. In *Proc. Joint Int'l Conf. and Symp. on Logic Programming*, pages 289–303, 1996.

[45] Judea Pearl. Embracing causality in default reasoning. *Artificial Intelligence*, 35:259–271, 1988.

[46] Teodor Przymusinski and Hudson Turner. Update by means of inference rules. *Journal of Logic Programming*, 30(2):125–143, 1997.

[47] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1,2):81–132, 1980.

[48] Erik Sandewall. Assessments of ramification methods that use static domain constraints. In *Principles of Knowledge Representation and Reasoning: Proc. of the Fifth Int'l Conference*, 1996.

[49] Michael Thielscher. Computing ramifications by postprocessing. In *Proc. of IJCAI-95*, pages 1994–2000, 1995.

[50] Michael Thielscher. Ramification and causality. *Artificial Intelligence*, 89(1–2):317–364, 1997.

[51] Hudson Turner. Representing actions in logic programs and default theories: A situation calculus approach. *Journal of Logic Programming*, 31(1–3):245–298, 1997.

[52] Richard Watson. A practical application of action theory to the space shuttle. In *Proc. of Practical Aspects of Declarative Languages*, 1999. To appear.

[53] Marianne Winslett. Reasoning about action using a possible models approach. In *Proc. of AAAI–88*, pages 89–93, 1988.